Identification and Characterization of Pathogenic Mutations in Neurodevelopmental Disorders Discovered by Next-generation Sequencing

by

Elizabeth Kathryn Ruzzo

University Program in Genetics and Genomics
Duke University

Date:_____

Approved:

_____
David B. Goldstein, Supervisor

_____
Michael A. Hauser, Chair

_____
Frederick S. Dietrich

_____
Rodney A. Radtke

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy
in the University Program in Genetics and Genomics
in the Graduate School
of Duke University

2014

ABSTRACT

Identification and Characterization of Pathogenic Mutations in Neurodevelopmental Disorders Discovered by Next-generation Sequencing

by

Elizabeth Kathryn Ruzzo

University Program in Genetics and Genomics
Duke University

Date:_____

Approved:

_____
David B. Goldstein, Supervisor

_____
Michael A. Hauser, Chair

_____
Frederick S. Dietrich

_____
Rodney A. Radtke

An abstract of a dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy
in the University Program in Genetics and Genomics
in the Graduate School
of Duke University

2014

Abstract

Neurodevelopmental disorders develop over time and are characterized by a wide variety of mental, behavioral, and physical phenotypes. The categorization of neurodevelopmental disorders encompasses a broad range of conditions including intellectual disability, autism spectrum disorder, attention deficit hyperactivity disorder, cerebral palsy, schizophrenia, bipolar disorder, and epilepsy, among others. Diagnostic classifications of neurodevelopmental disorders are complicated by comorbidities among these neurodevelopmental disorders, unidentified causal genes, and growing evidence of shared genetic risk factors.

We sought to identify the genetic underpinnings of a variety of neurodevelopmental disorders, with a particular emphasis on the epilepsies, by employing next-generation sequencing to thoroughly interrogate genetic variation in the human genome/exome. First, we investigated four families presenting with a seemingly identical and previously undescribed neurodevelopmental disorder characterized by congenital microcephaly, intellectual disability, progressive cerebral atrophy, and intractable seizures. These families all exhibited an apparent autosomal recessive pattern of inheritance. Second, we investigated a heterogeneous cohort of ~60 undiagnosed patients, the majority of whom suffered from severe neurodevelopmental disorders with a suspected genetic etiology. Third, we investigated 264 patients with epileptic

encephalopathies – severe childhood epilepsy disorders – looking specifically at infantile spasms and Lennox–Gastaut syndrome. Finally, we investigated ~40 large multiplex epilepsy families with complex phenotypic constellations and unclear modes of inheritance. The studied neurodevelopmental disorders exhibited a range of genetic complexity, from clear Mendelian disorders to common complex disorders, resulting in varying degrees of success in the identification of clearly causal genetic variants.

In the first project, we successfully identified the disease-causing gene. We show that recessive mutations in *ASNS* (encoding asparagine synthetase) are responsible for this previously undescribed neurodevelopmental disorder. We also characterized the causal mutations *in vitro* and studied Asns-deficient mice that mimicked aspects of the patient phenotype. This work describes ASNS deficiency as a novel neurodevelopmental disorder, identifies three distinct causal mutations in the ASNS gene, and indicates that asparagine synthesis is essential for the proper development and function of the brain.

In the second project, we exome sequenced 62 undiagnosed patients and their unaffected biological parents (trios). By analyzing all identified variants that were annotated as putatively functional and observed as a novel genotype in the probands (not observed in the unaffected parents or controls), we obtained a genetic diagnosis for 32% (20/62) of these patients. Additionally, we identify strong candidate variants in 31% (13/42) of the undiagnosed cases. We also present additional analysis methods for moving beyond traditional screens, e.g., considering only securely implicated genes, or

subjecting qualifying variants from any gene to two unique analysis approaches. This work adds to the growing evidence for the utility of diagnostic exome sequencing, increases patient sizes for rare neurodevelopmental disorders (enabling more detailed analyses of the phenotypic spectrum), and proposes novel analysis approaches which will likely become beneficial as the number of sequenced undiagnosed patients grows.

In the third project, we again employ a trio-based exome sequencing design to investigate the role of *de novo* mutations in two classical forms of epileptic encephalopathy. We find a significant excess of *de novo* mutations in the ~4,000 genes that are the most intolerant to functional genetic variation in the human population (P = 2.9×10^{-3}, likelihood analysis). We provide clear statistical evidence for two novel genes associated with epileptic encephalopathy – *GABRB3* and *ALG13*. Together with the 15 well-established epileptic encephalopathy genes, we statistically confirm the association of an additional nine putative epileptic encephalopathy genes. We show that only ~12% of epileptic encephalopathy patients in our cohort are explained by *de novo* mutations in one of these 24 genes, highlighting the extreme locus heterogeneity of the epileptic encephalopathies.

Finally, we investigated multiplex epilepsy families to uncover novel epilepsy susceptibility factors. Candidate variants emerging from sequencing within discovery families were further assessed by cosegregation testing, variant association testing in a case-control cohort, and gene-based resequencing in a cohort of additional multiplex

epilepsy families. Despite employing multiple approaches, we did not identify any clear genetic associations with epilepsy. This work has, however, identified a set of candidates that may include real risk factors for epilepsy; the most promising of these is the *MYCBP2* gene. This work emphasizes the extremely high locus and allelic heterogeneity of the epilepsies and demonstrates that very large sample sizes are needed to uncover novel genetic risk factors.

Collectively, this body of work has securely implicated three novel neurodevelopmental disease genes that inform the underlying pathology of these disorders. Furthermore, in the final three studies, this work has highlighted additional candidate variants and genes that may ultimately be validated as disease-causing as sample sizes increase.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my advisor, David B. Goldstein for providing me with incredible research project opportunities and for granting me so much independence in perusing this work. I have learned a great deal from him about how to run a successful lab and how to participate in scientific collaborations. I would also like to acknowledge the other members of my thesis committee, Michael A. Hauser, Frederick S. Dietrich, and Rodney A. Radtke, for their incredible support and advice throughout my graduate school career. I would like to thank Erin L. Heinzen for her mentorship. Erin and I worked together on many epilepsy projects and she was always willing to sit down and carefully think through a problem with me. I have much gratitude for Yong-hui Jiang and Debra L. Silver for helping me with my molecular biology experiments and for inviting me into their labs and lab meetings as one of their own students. I am incredibly grateful for all of these scientific mentors who have helped me become a creative and meticulous scientist.

I would like to thank all members of the Goldstein Laboratory/Center for Human Genome Variation, past and present, for their collaborative spirit, helpful discussions, technical assistance, support, and friendship. Especially, Liz Cirulli, Kimberly Pelak, Erin Heinzen, Molly Cook, Anna Need, Jessica Maia, Nicole Walley, Brian Krueger, and Curtis “Hubby” Gumbs.

In addition, I would like to acknowledge my close collaborators over the years: Ruth Ottman, Sam Berkovic, Ingrid Scheffer, Doron Lancet, Elon Pras, Bruria Ben Zeev, Danit Oz-Levi, Debra Silver, Helen Mao, Yong-hui Jiang, Andrea Pappas and all members of the EPIGEN and Epi4K consortia. I have learned much from each of you. I also acknowledge the funding sources which made my research possible: Jo Rae Wright Fellowship for Outstanding Women in Science, Predoctoral Research and Training Fellowship from The Epilepsy Foundation, Ray J. Tysor Graduate Fellowship, and award number 5RC2NS070344-02 and award number NS077364 from the National Institute of Neurological Disorders and Stroke.

I would not have made it through graduate school with out the love and support of my friends and family. I would like to thank my Durham Sisters: Andrea Pappas and Megan Martik, for their unconditional support, sharing in the ups and downs of graduate school, countless laughs, adventures, and true friendship. To all of the talented and amazing friends I have made along the way: Jennifer Doss, Ashley Trama, Dawn Kernagis, Ashley van Heteren, Lauren Jackson, Jamie Weyandt, Charlotte Agger, and many more. A special thank you to my "Sangha": Jeremy Lipkowitz, Jessie Uehling, Aaron Towers, and the Duke BMC, who supported my mental health and taught me how to take care of myself even in stressful or uncertain times. I feel truly fortunate to have you all in my life.

Finally, I would like to thank my parents, Barbara and Larry Ruzzo, for instilling in me the importance of education and for always supporting and believing in me. A special thank you to my Dad for always being willing to help proof read a document or debug my computer code and my Mom for knowing how to support me even from across the country. I would like to thank my big sister, Emily Ruzzo, for being the best big sister in the history of sisters. She always believed in me and is the reigning queen of pep talks. Without you three, I do not think I would have completed this journey.

1. Introduction[1]

1.1 The human genome

1.1.1 An overview

Deoxyribonucleic acid, or DNA, is the hereditary material in humans and almost all other living organisms. DNA is made up of four unique chemical bases, or nucleotides, including: adenine (A), cytosine (C), guanine (G) and thymine (T). This alphabet of nucleotides provides instructions for sequences of amino acids (3 nucleotide combinations code for each of 20 amino acids), which the body uses to build proteins – the workhorses of the cell. A segment of DNA that codes for a protein is called a gene. The central dogma of molecular biology states that a gene is transcribed into a messenger ribonucleic acid (mRNA) transcript, which is then translated into a protein. Before the final mRNA is translated into protein, the pre-mRNA transcript contains both introns and exons. Introns are removed from the transcript and the exons are the portion of the gene that gets directly translated into protein.

The human genome consists of three billion base pairs of DNA and ~20,000 genes. The proportion of the human genome that is protein-coding ("the exome") is very small, accounting for less than 2% of the genome. The remaining 98% of the genome,

---

[1] Portions of this chapter are being submitted for publication in a textbook for medical students entitled "Practical Epilepsy" to be published by Demos Medical Publishing.

noncoding DNA, is not used to encode proteins. We don't fully understand the function of all noncoding DNA; however, some of it encodes RNA molecules with important biological functions (an exception to the central dogma) and other regions have important roles in regulating the expression of genes (whether, and how much, of a specific gene product is available to a cell).

DNA usually exists as double-stranded DNA in which Watson-Crick base pairs (guanine-cytosine and adenine-thymine pairing via hydrogen bonds) create a regular helical structure with a sugar phosphate backbone. This double helix enables the cell to easily copy the DNA molecule during cell division, enabling precise replication of our genetic material. The DNA double helix is further organized into 23 chromosomes: 22 autosomes (1-22) and the sex chromosomes (X and Y). The human sex cells (female ova and male sperm) are haploid, meaning they contain a single copy of our genome (3 billion base pairs on 23 chromosomes). Importantly, the other ~50 trillion cells in our body are diploid, meaning they have two copies of the human genome – one inherited from our mother and one from our father. Each human diploid cell has 6 billion base pairs of DNA and without additional modification this DNA would be 2 meters in length. DNA is further condensed to form chromatin. DNA wraps around proteins called histones creating a series of nucleosomes (nine histone proteins + 166 base pairs of DNA) with intervening "linker DNA" of ~20 base pairs; when viewed with an electron

microscope this gives the appearance of beads on a string. Thus nucleosomes are the structural unit of chromatin and further coiling generates tightly compacted higher-order structures. Chromosomes are most highly compacted during metaphase (a phase in mitosis of the cell cycle) and metaphase chromosomes can be viewed under a light microscope.

Finally, in addition to the chromosomal genomic DNA (gDNA), which resides in the nucleus of the cell, humans also harbor a small amount of mitochondrial DNA (mtDNA). As suggested by the name, this DNA resides in the mitochondria of the cell. Mitochondria produce energy for the cell and humans have hundreds to thousands of mitochondria per cell. Mitochondrial DNA contains 37 genes, packaged in a single circular chromosome; mtDNA is maternally inherited.

1.1.2 Genetic variation

When we compare the human genome to that of our closest living relatives, the chimpanzees, we see differences in only ~1% of our genomes. When we compare the genomes of human individuals we are ~99.9% identical. So why are individuals so unique? One of the main reasons comes from genetic variation. Even that 0.1% difference means we have many different changes in our genomes. In our 3 billion base pair genome, we have ~3.5 million sites[1] (or loci) where there is a single base pair change with respect to the reference genome; these are known as single nucleotide

variants (SNVs). Some SNVs explain differences in our physical features, others are related to disease or drug responses, but the majority of SNVs have no known phenotypic consequences. In addition to SNVs, there are other types of genetic variation including small insertions or deletions (indels), copy number variants (CNVs) and larger structural variants (Figure 1).

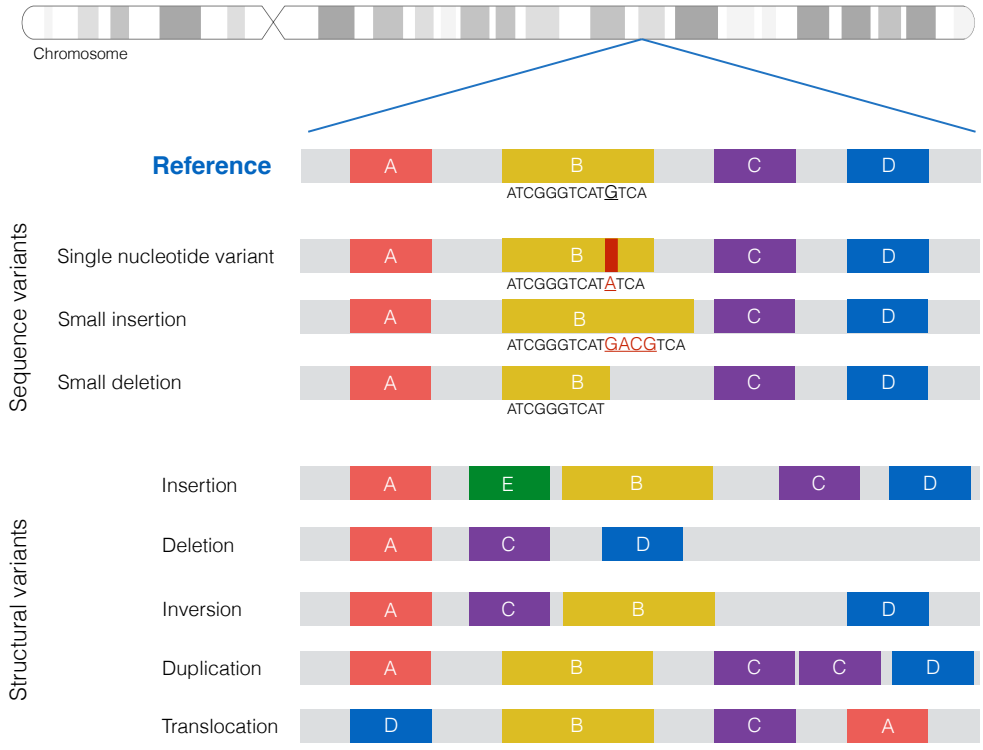


Figure 1. Types of variation in the human genome.

Each genetic variant is found at a certain frequency in the human population, with different subpopulations/ethnicities often having different frequencies for the same

variant (population stratification). Typically, this frequency is described by the minor allele frequency (MAF). The MAF for a variant locus is between 0 and 50% and reflects the proportion of alleles (in the population) that are the less frequent allele (the "variant" allele). If a genetic variant is variable within or between populations it is considered a genetic polymorphism (in contrast, genetic variants can be private). Therefore, a common single nucleotide variant is also known as a single nucleotide polymorphism (SNP). Genetic variants with different frequencies are detectable using different technologies and have unique implications for genetic analyses (Table 1).

Table 1. Frequencies of variants in the human population[2].

Variant Category	Minor allele frequency (MAF)	Implications for genetic analyses
Common	5-50%	Used in traditional GWA studies
Less Common	1-5%	Variants catalogued more recently and included in the newest GWA chips for association testing
Rare	Less than 1% but polymorphic in one or more major human populations	Detectable by NGS and amenable to analysis for case-control, family, or trio based studies
Private	Much less than 1% and found only in a single or a handful of analyzed samples and their immediate relatives	Difficult to gather statistical evidence except through co-segregation in families

[2]Table adapted from Cirulli *et al.* [2].

1.1.3 Copy number variation

Deletions, insertions, duplications, and complex rearrangements of large segments of genomic DNA are all forms of structural variation (Figure 1). Copy number variations (CNVs), are submicroscopic structural variants. CNVs are similar to single nucleotide polymorphisms (SNPs) in that they occur throughout the human genome and confer inter-individual genetic variation. If a CNV is present in >1% of the human population it is called a copy number polymorphism (CNP). A SNP alters a single nucleotide pair, whereas a CNV alters anywhere from one thousand base pairs (one kilobase, Kb) to several million base pairs (megabases, Mb) of DNA. Recurrent CNVs are copy number variants where the end-points (beginning and end of a duplication or deletion) are limited to a narrow genomic region with extensive homology. The homology of these regions makes it more likely for these CNV events to occur and thus, these CNVs are found in multiple individuals. In contrast, non-recurrent CNVs have very limited homology at their end-points and thus the same end-points are rarely observed in the human population.

While there are many more SNPs than CNVs in the human genome, CNVs impact a larger proportion of the genome with at least 10% of the genome being subject to copy number variation[3]. CNVs are found in both gene-rich and gene-poor regions.

CNVs are a major genetic component of phenotypic diversity, they can be nonpathogenic and are observed in "healthy" individuals with no clinical diagnosis.

1.2 Identifying genetic determinants of human phenotypes

1.2.1 Linkage studies

Linkage analysis is a traditional method for identifying a candidate region for a gene associated with a given disorder. Linkage analysis makes use of crossovers, or recombination events, that occur naturally during meiosis. Linkage analysis is conducted in either a single large pedigree or across multiple smaller pedigrees with identical phenotypes. Polymorphic genetic markers (e.g., SNPs or microsatellites) distributed throughout the genome, are then genotyped in all individuals of the family. If the affected individuals in a pedigree nearly always inherit a genetic marker, then the disease gene and the marker are likely to be close together on the chromosome. Each marker can then be tested for segregation with the disease phenotype ("co-segregation") and the disease can be statistically "linked" to a specific region of the genome. In linkage analysis, a logarithm of odds (LOD) score compares the likelihood of obtaining the observed data if the tested loci are indeed linked, to the likelihood of observing the same data purely by chance. If two genetic markers are on different chromosomes then there is a 50-50 chance that they are inherited together and thus they are "unlinked". In contrast, two genetic markers that are close together on the same chromosome have a

high chance of being inherited together ("linked") since it is less likely they will be separated by a meiotic recombination event. Thus in linkage analysis, a "linked" marker has a high LOD score indicating that very few meiotic recombination events have occurred between this marker and the disease gene, thus highlighting a chromosomal region of interest.

Once the candidate genetic region is mapped, the position of the disease-associated gene can be located on the chromosome through isolation of partially overlapping DNA segments that progress along the chromosome toward the disease gene, a technique referred to as positional cloning. Today, fine mapping can instead be achieved by referring to the human reference genome and subsequently sequencing (all genes or just candidate genes) within this linkage "peak" to directly identify the causal allele. Positional cloning in conjunction with family based linkage mapping and analysis can work for gene identification even without knowledge of the biochemical nature of a disease.

1.2.2 Genetic association studies

Association studies seek to determine if two things occur together more often than expected by chance. In a classical genetic association study, these "things" are (1) a single-locus allele or genotype and (2) a phenotype (disease cases vs. healthy controls). In other words, genetic association studies can identify genetic variants that are

associated with a disease or trait. If a genetic association increases susceptibility to a given disease, then the associated genetic variant will be seen more often than expected by chance in diseased individuals. While this sounds simple enough, there are many critical methodology issues to consider when conducting association studies; and true genetic associations are the result of carefully conducted studies involving large cohorts, replication cohorts, and statistically robust methods[4,5].

1.2.2.1 Candidate gene association studies

Before genome-wide approaches to association studies were readily accessible, many scientists conducted association studies based on a candidate gene approach. A candidate gene may have been selected for any number of reasons including biological plausibility for the phenotype of interest. The general approach of these studies was to use common SNPs as genetic markers within a given candidate gene and compare the frequency of alleles in affected cases to those in unaffected controls. These studies frequently resulted in multiple conflicting and nonreplicable results; one of the main reasons for this inconsistency is the inability to accurately account for population stratification in these analyses [4,6].

1.2.2.2 Genome wide association studies

Across many diverse disorders, candidate gene studies failed to identify definitive common genetic risk factors, and thus scientists recognized that a genome-

wide approach was needed to obtain unbiased assessment of markers throughout the genome. Two aspects of human population genetics indicated that an "indirect approach", of assaying a set of genetic markers – even if the markers themselves had unknown functional effects – would still capture most of the common patterns of variation in the human genome and thus detect regions of the genome associated with a phenotype[7]. The first relevant human population genetics observation was that ~90% of the genetic variants among individuals are common variants (MAF>5%)[8]. The second was that the majority of common variants arose from a single mutation event that occurred on an ancestral chromosome and thus these SNPs frequently occur in combination with nearby variants.

Around this same time, two critical large-scale efforts towards understanding variation in the human genome made this vision realistic: the human genome project (draft announced in 2000) and the International HapMap project (first data release in 2003). Additionally, microarray technology enabled high-throughput genotyping of hundreds of thousands of SNPs. This led to the advent of Genome-Wide Association (GWA) studies.

GWA studies use dense arrays of genetic markers, typically SNPs, to survey a large proportion of common variants in the human genome. SNPs are either genotyped directly or indirectly through linkage disequilibrium. Linkage disequilibrium (LD) is the

nonrandom association between alleles at different loci; these loci are often in close physical proximity since the likelihood of recombination between two sites increases with the distance between them. GWA studies attempt to identify associations between genotype frequency and trait status (affected patient vs. healthy individual). The effect size describes the increased population risk for a given trait that is conferred by a given genetic variant. GWA studies have had some success in common diseases, however the associated SNPs typically have modest effect sizes and even when all associated variants are considered collectively, they still explain only a small fraction of known heritability and thus have limited translational potential in the clinic. A current catalog of published GWA studies can be found at the National Human Genome Research Institute's (NHGRI) website (https://www.genome.gov/26525384).

1.2.3 Next-generation sequencing

1.2.3.1 An overview

The Human Genome Project sought to sequence the entire human genome, in other words, to determine the exact order of the base pairs in the human genome. This project used DNA from multiple anonymous volunteers and took over ten years to complete. Next-generation sequencing (NGS), also known as massively parallel sequencing (MPS), has revolutionized the cost and speed with which a human genome can be sequenced – with genomes now being sequenced within a week. While the

technical details vary by the sequencing platform used, these high-throughput sequencing approaches generate millions of short sequence reads in parallel. These short sequence reads can then be aligned to the human reference genome (generated by the Human Genome Project). A computer algorithm is then used to perform "variant calling", which results in the identification of all alleles in the newly sequenced genome that differ from the reference genome, including SNVs and indels. Additional algorithms have also been developed to identify structural variants from whole genome and exome sequence data[9-11]. In exome-sequencing, an additional step is included prior to sequencing that targets and captures only the exonic and flanking intronic base pairs. Exome-sequencing is popular for two main reasons: a) the majority of known disease-causing mutations are in protein coding regions of the genome and b) exome-sequencing targets ~2% of the human genome and thus is cheaper and faster than whole-genome sequencing. In contrast, whole-genome sequence data can be used to identify non-coding variants whose function we will likely understand more fully in the coming years with the advent of projects like the ENCyclopedia Of DNA Elements (ENCODE)[12].

1.2.3.2 Next-generation sequencing study designs

Next-generation sequencing (NGS) facilitates a thorough interrogation of nearly all the genetic variants in the genome, including very rare variants not directly

interrogated using GWA methods. These genetic variants must be prioritized differently for different diseases and study designs. At the broadest level a number of factors should be considered, including: the mode of inheritance (e.g., if recessive, focus on homozygous variants), the frequency of the disorder (e.g., if the disorder is rare, the causal variant(s) will be very rare or absent in controls), and predicted deleterious nature of the variant itself. Finally, a variant-based approach could be used if the hypothesis is that causal variants will have a relative large effect and will be present in multiple cases. In contrast, if the causal variants are individually very rare in the case population but are hypothesized to lie within the same gene(s), then gene-based approaches should be used.

Research has already established that in Mendelian disease whole genome or exome sequencing of even just a small number of cases can readily identify the causal variants as those that are shared amongst a small number of unrelated affected individuals and rare in the general population[13-15]. In contrast, a number of different NGS study designs can be considered for optimizing discovery of disease-associated variants for complex diseases. Firstly, a classical case-control study design can be used in which a large number of case samples and ethnically matched controls are sequenced to detect variants (or genes with qualifying variants) that are enriched in the case population. The main disadvantage to this approach is that very large sample sizes are

needed, which is still cost prohibitive, to perform sufficiently powered whole-genome association studies. However, power can be increased by either: restricting the tested variants based on *a priori* predictions of the functional impact of the variants or by sequencing individuals on one or both extreme ends of a phenotypic distribution[16].

Secondly, a trio-based design can be used in which the healthy biological parents and affected children are sequenced and newly formed genotypes (e.g., *de novo*, newly homozygous) are identified in the child. The average human germline mutation rate is approximately 1.2×10^{-8} per base pair per generation, which equates to roughly 70 *de novo* mutations per generation. Therefore, each individual is expected to have only ~1 exonic *de novo* mutation. *De novo* mutations, especially those predicted to damage an encoded protein, can be disease causing and are increasingly surveyable with NGS trio studies.

Thirdly, if families with multiple relatives affected with a complex disease exist, then family based studies can be utilized. While any number of relatives could be selected for sequencing, one cost effective strategy would be to sequence distantly related diseased individuals to minimize the number of variants shared by chance while still enriching for any shared risk variant(s). These shared variants can then be tested for cosegregation of the variant with affectation status in the entire family pedigree.

1.3 Epilepsy

Epilepsy is one of the most common neurological disorders, affecting ~3% of the human population at some period during life, with children and the elderly having the highest incidences[17]. Epilepsy is characterized by recurrent (≥2), unprovoked seizures. A seizure is the clinical manifestation of abnormal synchronization and excessive electrical discharge in the brain; in the clinic, an electroencephalograph (EEG) is used to record the electrical activity produced by the firing of neurons in the brain.

The treatment of epilepsy is extremely challenging. Despite the availability of many anticonvulsant medications, over 30% of epilepsy patients lack adequate seizure control[18]. Even in individuals who obtain some level of seizure control, the control is often achieved by "trial and error" to find the appropriate medication, dose, or combination of medications. In addition to the lack of clear treatment protocols, there are many adverse drug responses associated with the existing anticonvulsants. Fortunately, at least a small proportion of epilepsy patients obtain seizure freedom via brain surgery (only an option for patients where a single non-vital portion of the brain (focal epilepsies) is responsible for seizures), single drug treatment ("monotherapy"), or the ketogenic diet. One avenue for improving patient treatment is to discover genes that influence epilepsy risk, thus providing additional drug targets for drug development.

Epilepsy is a heterogeneous disorder made up of many (over 30[19]) unique epilepsy syndromes and nonsyndromic cases. At the most basic level, epilepsy can be divided into two broad categories based on the primary localization of abnormal brain activity: generalized or focal. Generalized epileptic seizures do not necessarily include the entire cortex, but from the point of origin they rapidly engage bilaterally distributed networks of neurons[19]. While individual seizure onsets can appear localized, the location is not necessarily consistent from one seizure to another. In contrast, focal seizures originate within networks limited to one hemisphere, and they may impact only a small, localized network or they may have a wider distribution[19]. There is also phenotypic diversity within these two broad seizure types, governed by the region and extent of the brain that participates in the abnormal electrical activity. Some individuals experience strange sensations or behaviors while others experience quintessential convulsions and even loss of consciousness.

To diagnose a specific epilepsy syndrome, the type of seizure is considered together with any number of other factors such as age of onset, specific region of the brain involved, events that provoke a seizure, cause of the seizures, or electroencephalography (EEG) patterns.

It is also worth noting that a number of individuals develop what is known as acquired epilepsy, implying that some type of insult or provocation (central nervous

system infection or brain trauma) preceded the onset of epilepsy. While there is some evidence that there is still a genetic predisposition to developing acquired epilepsy, these individuals are not the primary focus of this research.

Despite the diverse etiologies of epilepsy, it is highly heritable and genetics play an important etiological role. A number of different lines of epidemiological evidence support this claim including: higher concordance rates in monozygotic (49%) compared to dizygotic (16%) twins[20] and increased risk in first-degree relatives of probands[21-23].

Epilepsy can be broadly classified by the suspected etiology: structural–metabolic[19] (previously known as symptomatic) cases are explained by brain malformations, tumors, strokes, or other detectable phenotypes, unknown[19] (previously known as cryptogenic) cases are suspected to be explained by a clinical neurological abnormality but the cause has not been identified, and finally, genetic[19] (previously known as idiopathic) cases have no obvious cause, but are presumably genetic. Approximately 30% of all epilepsies are considered to be idiopathic and thus likely have a genetic basis[24]. The International League Against Epilepsy (ILAE) Commission on Classification and Terminology now refers to idiopathic generalized epilepsy as genetic generalized epilepsy (GGE)[19]. Genetic generalized epilepsies (GGE) have an especially strong genetic component, with ~80% concordance for

monozygotic twins[25]; focal epilepsies, in contrast, have a monozygotic twin concordance of 36%[25]. To complicate matters further, there is evidence that some structural-metabolic cases may also have underlying genetic factors that interact with environmental factors to increase disease susceptibility[26].

Diseases explained by simple monogenic Mendelian patterns of inheritance are rare compared to complex disorders that violate Mendelian inheritance. Complex disorders are caused by interplay between genetic and environmental factors. Epilepsy is a complex disorder, however a very small proportion of epilepsy cases, roughly 1%, show genetic transmission in a Mendelian pattern. The majority of epilepsy cases are either sporadic, with no known family history, or cluster in families that do not show a clear-cut pattern of inheritance.

1.3.1 An overview of discovery genetics in epilepsy

The initial efforts to identify genes influencing epilepsy risk came from linkage studies in rare epilepsy families with Mendelian inheritance patterns. These familial linkage studies identified over 20 "epilepsy genes"[27]. However, mutations in these genes only account for an estimated 1% of epilepsy cases. This highlighted the difficulties associated with genetic discovery in a clinically and genetically heterogeneous disorder, such as epilepsy.

The next efforts came from candidate gene, and later, genome-wide association (GWA) studies. Association studies were promising because, unlike linkage analyses that rely on acquisition of multiplex families, they could be conducted in case-control populations by comparing the frequency distribution of one or more variants. Unfortunately, candidate gene association studies were underpowered and no convincing susceptibility genes were identified using this approach[4]. Three GWA studies completed to date provided only modest evidence for additional loci and replication of these signals is needed to prove their true association with epilepsy.

Despite these intensive research efforts, scientists cannot explain the genetic basis of epilepsy in the vast majority of patients. The failure of association studies to discover common disease-associated variants lends credence to a rare variant-common disease model for epilepsy[28,29]. Recently, the role of rare variation in human disease has become increasingly clear, especially the role of rare copy number variations (CNVs) in neuropsychiatric disorders and epilepsy in particular[30-35]. Current efforts are focused on the use of next-generation sequencing to systematically explore the role of rare variation in epilepsy.

Over one hundred genes have been associated with epilepsy (http://www.epigad.org; February 2014), however the number of securely established genes is closer to 50 (Table 2).

1.3.2 Early epilepsy genetics and known epilepsy genes

Epilepsies appearing in multiplex pedigrees with Mendelian patterns of inheritance facilitated the first genetic discoveries in epilepsy and informed our understanding of the underlying biology of seizures. In 1995, the first epilepsy associated gene was identified in families with autosomal dominant nocturnal frontal lobe epilepsy (ADNFLE)[36]. By studying one large ADNFLE family, a region of interest was initially identified on chromosome 20 and subsequently this candidate region was narrowed to a single causal missense mutation in *CHRNA4*, encoding the cholinergic receptor, nicotinic, alpha 4.

Such family based linkage mapping studies, in the late 1990s, lead to the identification a number of epilepsy genes without prior knowledge of the biochemical nature of a disease. The earliest epilepsy genes fell into several main categories – voltage-gated or ligand-gated ion channels, namely subunits of acetylcholine receptors (*CHRNA2*, *CHRNA4*, and *CHRNB2*), subunits of sodium channels (*SCN1A*, *SCN1B*, and *SCN2A*), subunits of potassium channels (*KCNQ2* and *KCNQ3*), and subunits of Gamma-Aminobutyric Acid (GABA) receptors (*GABRA1* and *GABRG2*). This led to the perception that epilepsy was primarily a "channelopathy", resulting from disruption of normal electrical transmission between neurons.

While these classes of genes are still central to epileptogenesis, additional research has identified many new classes of genes (Table 2) highlighting the need to reconsider our narrow view of the properties of epilepsy genes. A diversity of examples exist, including: i) leucine-rich, glioma-inactivated 1 gene (*LGI1*)[37], a synaptic protein which may also regulate voltage-gated potassium channels ii) Dishevelled, Egl-10 and Pleckstrin (DEP) domain–containing protein 5 (*DEPDC5*) of unknown function[38,39] iii) Proline-Rich Transmembrane Protein 2 (*PRRT2*), also of unknown function[40] and iv) Chromodomain Helicase DNA Binding Protein 2 (*CHD2*)[41], a chromatin remodeling protein.

In total, linkage studies in these Mendelian epilepsy families identified over 20 "epilepsy genes"[27] (Table 2). Collectively, however, these genes explain only an estimated 1% of epilepsy cases. The majority of unsolved epilepsy cases are considered complex epilepsies, which will be explained by some combination of gene-gene or gene-environment interactions. It is also likely that the extent of genetic heterogeneity in the complex epilepsies will be greater than that of the Mendelian epilepsies. This would mean that many different genes would each explain only a very small proportion of epilepsy cases, making them evasive to current genetic approaches.

Table 2. Epilepsy genes and their associated syndromes.

This list was generated from searching the OMIM gene map for phenotypes including the words "epilepsy", "epileptic", or "seizure" and filtered to exclude associations with weak supporting evidence. This database is updated regularly; this table is based on findings as of January 21st, 2014. Additionally, several recent discoveries have not been entered into OMIM yet but are included: *ALG13*[42], *GABRB3*[42], and *SYNGAP1*[41]. The most well-accepted and validated genes with a major effect on susceptibility Mendelian idiopathic epilepsies are listed in bold[27]. AD: autosomal dominant; AR: autosomal recessive; XD: X-linked dominant; XR: X-linked recessive; X: X-linked female only (*PCDH19* is associated with female-restricted epilepsy) or X-linked but mode of inheritance unclear from OMIM reports.

Gene Symbol	Gene Description	Cytogenetic location	Associated Epilepsy Phenotype(s)	MIM number(s)	Mode of Inheritance
ARX	Aristaless-related homeobox, X-linked	Xp21.3	Epileptic encephalopathy, early infantile, 1	308350	XR
CDKL5	Cyclin-dependent kinase-like 5	Xp22.13	Epileptic encephalopathy, early infantile, 2	300672	XD
CHRNA2	Cholinergic receptor, nicotinic, alpha polypeptide-2	8p21.2	Epilepsy, nocturnal frontal lobe, type 4	610353	AD
CHRNA4	Cholinergic receptor, nicotinic, alpha polypeptide-4	20q13.33	Epilepsy, nocturnal frontal lobe, 1	600513	AD
CHRNB2	Cholinergic receptor, nicotinic, beta polypeptide-2	1q21.3	Epilepsy, nocturnal frontal lobe, 3	605375	AD
GABRG2	Gamma-aminobutyric acid (GABA) A receptor, gamma-2	5q34	Epilepsy, generalized, with febrile seizures plus, type 3; Febrile seizures, familial, 8	611277; 611277	AD; AD
KCNMA1	Potassium large conductance calcium-activated channel, subfamily M, alpha member 1 (slowpoke, Drosophila, homolog of)	10q22.3	Generalized epilepsy and paroxysmal dyskinesia	609446	AD
KCNQ2	Potassium voltage-gated channel, KQT-like subfamily, member 2	20q13.33	Seizures, benign neonatal, 1; Epileptic encephalopathy, early infantile, 7	121200; 613720	AD; AD
KCNQ3	Potassium voltage-gated channel, KQT-like subfamily, member 3	8q24.22	Seizures, benign neonatal, type 2	121201	AD
LGI1	Leucine-rich gene, glioma-inactivated, 1	10q23.33	Epilepsy, familial temporal lobe, 1	600512	AD
PCDH19	Protocadherin 19	Xq22.1	Epileptic encephalopathy, early infantile, 9	300088	X
SCN1A	Sodium channel, voltage-gated, type I, alpha polypeptide	2q24.3	Epilepsy, generalized, with febrile seizures plus, type 2	604403	AD
SCN1B	Sodium channel, voltage-gated, type I, beta polypeptide	19q13.12	Epilepsy, generalized, with febrile seizures plus, type 1	604233	AD
SCN2A	Sodium channel, voltage-gated, type II, alpha subunit	2q24.3	Seizures, benign familial infantile, 3; Epileptic encephalopathy, early infantile, 11	607745; 613721	AD; AD
SLC2A1	Solute carrier family 2 (facilitated glucose transporter), member 1	1p34.2	GLUT1 deficiency syndrome 1; GLUT1 deficiency syndrome 2	606777; 612126	AD ; AD
STXBP1	Syntaxin-binding protein 1	9q34.11	Epileptic encephalopathy, early infantile, 4	612164	AD

Gene Symbol	Gene Description	Cytogenetic location	Associated Epilepsy Phenotype(s)	MIM number(s)	Mode of Inheritance
ALDH7A1	Aldehyde dehydrogenase 7 family, member A1	5q23.2	Epilepsy, pyridoxine-dependent	266100	AR
ALG13	Alg13, S. cerevisiae, homolog of	Xq23	Epileptic encephalopathy (not in OMIM yet)	N/A	XD
ARHGEF9	Rho guanine nucleotide exchange factor 9	Xq11.1- q11.2	Epileptic encephalopathy, early infantile, 8	300607	XR
ASAH1	N-acylsphingosine amidohydrolase (acid ceramidase) 1	8p22	Spinal muscular atrophy with progressive myoclonic epilepsy	159950	AR
CHD2	Chromodomain helicase DNA binding protein-2	15q26.1	Epileptic encephalopathy, childhood-onset	615369	AD
CLN8	CLN8 gene	8p23.3	Northern epilepsy variant	610003	AR
CNTNAP2	Contactin-associated protein-like 2	7q35-q36	Cortical dysplasia-focal epilepsy syndrome	610042	AR
CPA6	Carboxypeptidase A6	8q13.2	Epilepsy, familial temporal lobe, 5; Febrile seizures, familial, 11	614417; 614418	AD ; AR
CSTB	Cystatin B (stefin B)	21q22.3	Epilepsy, progressive myoclonic 1A (Unverricht and Lundborg)	254800	AR
DEPDC5	DEP domain-containing protein 5	22q12.2-q12.3	Epilepsy, familial focal, with variable foci	604364	AD
EPM2A	Laforin	6q24.3	Epilepsy, progressive myoclonic 2A (Lafora)	254780	AR
GABRB3	Gamma-aminobutyric acid (GABA) A receptor, beta-3	15q12	Epileptic encephalopathy (not in OMIM yet)	N/A	AD
GNAO1	Guanine nucleotide-binding protein (G protein), alpha-activating activity	16q12.2	Epileptic encephalopathy, early infantile, 17	615473	AD
GOSR2	Golgi snap receptor complex member 2	17q21.32	Epilepsy, progressive myoclonic 6	614018	AR
GRIN2A	Glutamate receptor, ionotropic, N-methyl D-aspartate 2A	16p13.2	Epilepsy, focal, with speech disorder and with or without mental retardation	245570	AD
IER3IP1	Immediate-early response 3-interacting protein 1	18q21.1	Microcephaly, epilepsy, and diabetes syndrome	614231	AR
KCNT1	Potassium channel, subfamily T, member 1	9q34.3	Epileptic encephalopathy, early infantile, 14; Epilepsy, nocturnal frontal lobe, 5	614959; 615005	AD ; AD
KCTD7	Potassium channel tetramerization domain containing 7	7q11.21	Epilepsy, progressive myoclonic 3, with or without intracellular inclusions	611726	AR
MEF2C	MADS box transcription enhancer factor 2, polypeptide C (myocyte enhancer factor 2C)	5q14.3	Mental retardation, stereotypic movements, epilepsy, and/or cerebral malformations	613443	AD
NHLRC1	NHL repeat-containing 1 gene (malin)	6p22.3	Epilepsy, progressive myoclonic 2B (Lafora)	254780	AR
PLCB1	Phospholipase C, beta-1	20p12.3	Epileptic encephalopathy, early infantile, 12	613722	AR
PNKP	Polynucleotide kinase 3' phosphatase	19q13.33	Epileptic encephalopathy, early infantile, 10	613402	AR
PRICKLE1	Prickle-like 1	12q12	Epilepsy, progressive myoclonic 1B	612437	AR
PRICKLE2	Prickle-like 2	3p14.1	Epilepsy, progressive myoclonic 5	613832	AD
PRRT2	Proline-rich transmembrane protein 2	16p11.2	Seizures, benign familial infantile, 2; Convulsions, familial infantile, with paroxysmal choreoathetosis	605751; 602066	AD ; AD
SCARB2	Scavenger receptor class B, member 2	4q21.1	Epilepsy, progressive myoclonic 4, with or without renal failure	254900	AR

Gene Symbol	Gene Description	Cytogenetic location	Associated Epilepsy Phenotype(s)	MIM number(s)	Mode of Inheritance
SCN8A	Sodium channel, voltage gated, type VIII, alpha polypeptide	12q13.13	Epileptic encephalopathy, early infantile, 13	614558	AD
SCN9A	Sodium channel, voltage-gated, type IX, alpha subunit	2q24.3	Febrile seizures, familial, 3B; Epilepsy, generalized, with febrile seizures plus, type 7	613863; 613863	AD ; AD
SIAT9	Sialyltransferase 9	2p11.2	Amish infantile epilepsy syndrome	609056	AR
SLC25A22	Solute carrier family 25 (mitochondrial carrier, glutamate), member 22	11p15.5	Epileptic encephalopathy, early infantile, 3	609304	AR
SNIP1	SMAD nuclear interacting protein 1	1p34.3	Psychomotor retardation, epilepsy, and craniofacial dysmorphism	614501	AR
SPTAN1	Spectrin, alpha, nonerythrocytic-1 (alpha-fodrin)	9q34.11	Epileptic encephalopathy, early infantile, 5	613477	AD
SRPX2	SUSHI repeat-containing protein, X-linked, 2	Xq22.1	Rolandic epilepsy, mental retardation, and speech dyspraxia	300643	X
ST3GAL3	ST3 beta-galactoside alpha-2,3-sialyltransferase 3	1p34.1	Mental retardation, autosomal recessive 12; Epileptic encephalopathy, early infantile, 15	611090; 615006	AR ; AR
STRADA	STE20-related kinase adaptor alpha	17q23.3	Polyhydramnios, megalencephaly, and symptomatic epilepsy	611087	AR
SYN1	Synapsin I	Xp11.23	Epilepsy, X-linked, with variable learning disabilities and behavior disorders	300491	X
SYNGAP1	Synaptic Ras GTPase activating protein 1	6p21.32	Epileptic encephalopathy (not in OMIM yet)	N/A	AD
SZT2	Seizure threshold 2, mouse, homolog of	1p34.2	Epileptic encephalopathy, early infantile, 18	615476	AR
TBC1D24	TBC1 domain family, member 24	16p13.3	Myoclonic epilepsy, infantile, familial; Epileptic encephalopathy, early infantile, 16	605021; 615338	AR

1.3.3 Locus Heterogeneity & Variable Expressivity

Epilepsy shows extreme genetic heterogeneity. Locus heterogeneity is evident from the relatively large number of already established epilepsy genes. A single epilepsy syndrome may be caused by mutations in one gene in family A and caused by mutations in a different gene in family B. For example, Genetic Epilepsy with Febrile Seizures Plus (GEFS+) can be caused by mutations in *SCN1A*, *SCN2A*, *SCN1B*, or *GABRG2*. Variable expressivity is also observed in epilepsy; this is when mutations in a single gene can produce different epilepsy phenotypes in different individuals. For example, mutations in *SCN1A* can cause GEFS+ or Dravet syndrome. Until we have fully characterized all genotype-phenotype relationships, it will be difficult to understand the shared and distinct genetic influences on different epilepsy syndromes.

1.3.4 Syndromes with epilepsy as a feature

There are two broad categories of genes associated with epilepsy: those discovered in primary epilepsy syndromes and those discovered in syndromes with epilepsy as a feature (e.g., brain development disorders). Genes in the latter category have also primarily been identified by linkage analyses. Both categories still inform the biological mechanisms of epileptogenesis and provide possible therapeutic targets.

Clinically, the presentation of these syndromic and nonsyndromic phenotypes often confounds obtaining a clear diagnosis. Additionally, determining the genetic basis

of a patient's epilepsy may help guide treatment and inform counseling of recurrence risks. To aid in genetic diagnoses for epilepsy, a panel of 265 genes that are "most relevant" to epilepsy have been recommended for genetic testing[43]. This includes 18 phenotypic groupings or subpanels (Figure 2). Sequencing these 265 genes in 33 epileptic patients, resulted in the identification of a presumably causal variant in 48% of patients (n=16)[43]. In addition to acting as a diagnostic tool, use of this next-generation sequencing panel of 265 genes will also uncover the phenotypic heterogeneity associated with the less frequently mutated genes.

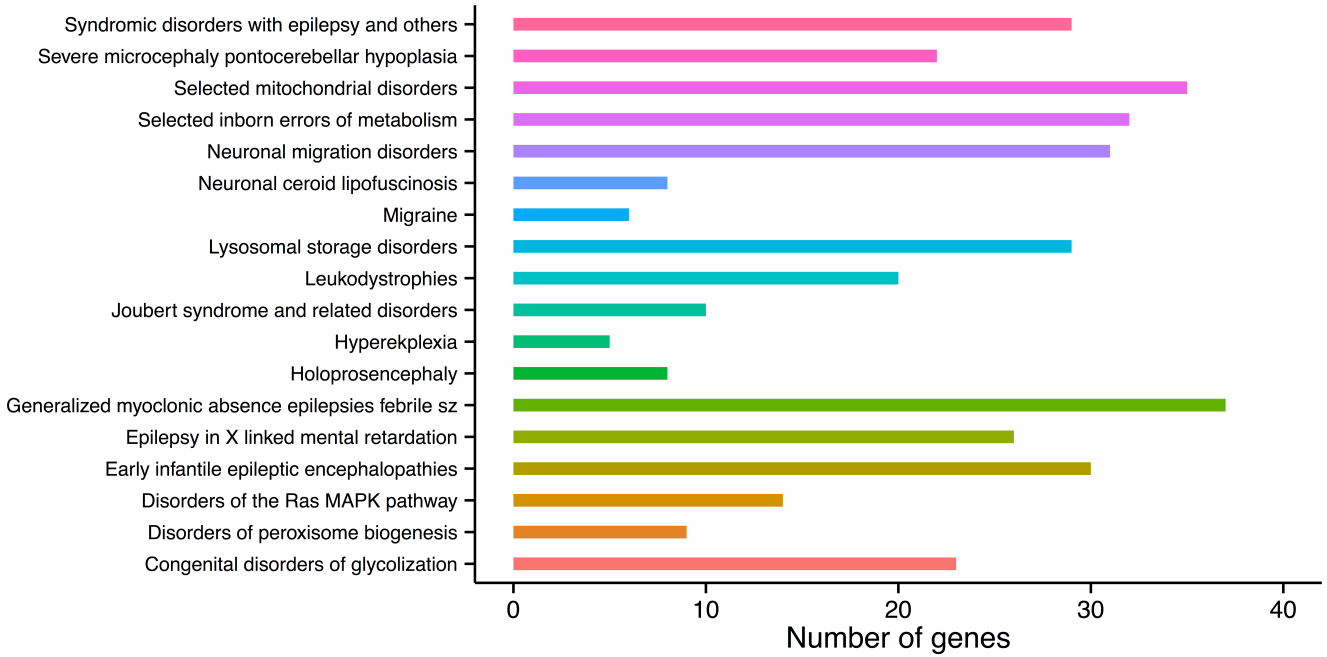


Figure 2. A bar graph of the recommended 364 genes for NGS diagnosis in epilepsy patients categorized by phenotypic grouping/subpanel[43].

This includes 280 unique genes and 44 genes in more than one phenotypic grouping (38 genes in two groups and 6 genes in three groups). *Note: original publication*[43] *says 365 genes are included on the panel but only 364 genes are listed in the original publication (324 unique genes).*

1.3.5 Genetic association studies in epilepsy

1.3.5.1 Candidate gene association studies in epilepsy

In the late 1990s and early 2000s, many genetic association studies were conducted for both focal and generalized epilepsies; in fact, over 50 studies were conducted involving hundreds of genes[4,6]. These studies resulted in multiple conflicting and nonreplicable results, and ultimately failed to identify any definitive common genetic risk factors for epilepsy[4,6].

1.3.5.2 Genome wide association studies in epilepsy

To date, three GWA studies have been conducted in epilepsy. The first study examined focal epilepsy patients of European ancestry[44]. This was a phenotypically heterogeneous cohort in that all focal epilepsies, regardless of syndrome or possible structural-metabolic causes, were included. This resulted in a cohort of ~3,500 cases and ~7,000 controls with no history of seizures. It is now widely-accepted that the threshold for genome-wide significance in association studies is 5×10^{-8} [5]; when correcting for the 528,745 SNPs genotyped in this study, the threshold required to achieve significance was 9.46×10^{-8}. However, the top SNP in this study had a p-value of 3.34×10^{-7} and thus no SNPs were found to be significantly associated with the focal epilepsies[44]. A second GWA study was also conducted in focal epilepsy patients; these patients were of Chinese ancestry[45]. This GWA study was divided into two stages: a discovery stage

(~500 cases (structural–metabolic focal) vs. ~3000 controls) and a replication stage (~600 cases (structural–metabolic or unknown focal) and ~500 controls). The initial discovery stage did not detect any variants of genome-wide significance. Next, they followed up a subset of SNPs with the lowest p-values (selected based on significance and regional LD structure) in the discovery stage by analyzing only these 80 SNPs in the replication stage and found one SNP that surpassed the threshold for genome-wide significance. This SNP resides on 1q32.1 in the *CAMSAP1L1* gene; this gene encodes a cytoskeletal protein whose biological connection to epilepsy is unclear. It is unclear if this finding is only relevant in this ethnic population and external replication is needed to prove the association with epilepsy. Finally, a third study was conducted in genetic generalized epilepsy (GGE) patients of European ancestry[46]. Again, this GWA study was conducted in two stages: a discovery stage (~1500 GGE cases vs. ~2400 controls) and a replication stage with two independent cohorts (~600 parent-offspring trios where the unaffected parents were treated as controls, and an additional ~900 GGE cases and ~900 controls). No SNPs reached genome-wide significance in the discovery stage. For the replication stage, they selected a subset of SNPs with the lowest p-values in the discovery stage (selected based on significance and regional LD structure), resulting in the analysis of ~20 SNPs in the replication stage. Again, no SNPs reached genome-wide significance. Finally, they performed a combined analysis with stage 1 and stage 2

samples and despite finding no associations of genome-wide significance, they highlight a SNP in the 5′ untranslated region (5′-UTR) of *SCN1A*. *SCN1A* has the largest number of known epilepsy-associated mutations[47] and thus this low signal (rs11890028, $P_{meta} = 4.0 \times 10^{-6}$) is likely due to this SNP being in LD with rare causal mutations in *SCN1A*. After separating the samples into two syndromic subgroups, genetic absence epilepsies (GAEs) and juvenile myoclonic epilepsy (JME), to look for syndrome-related variants, they also found weak non-genome-wide significant signals but larger samples sizes and replication will also be needed to prove these associations.

Taken all together, these three GWA studies provided only modest evidence for additional loci and replication of these signals is needed to prove their true association with epilepsy. This is likely due to a number of different factors. The first is simply that epilepsy is a highly heterogeneous disorder and thus obtaining large cohorts for well-powered associations studies requires well-phenotyped and phenotypically homogeneous samples, in very large numbers. Secondly, it is possible that the true causal variants exist at lower frequencies in the population and thus multiple rare causal variants of large effect may be driving the diluted signals observed when assaying common variants; if so, direct identification of the causal variants may provide a better approach[48].

1.3.6 Copy number variation in epilepsy

1.3.6.1 The role of copy number variants in epilepsy

The failure of association studies to discover common variants with a clear effect in epilepsy suggested that rare variants (found in <1% of the population) might be underlying the etiology of epilepsy. Further support of this hypothesis came from a number of studies in other neurological disorders, where strong evidence emerged for rare CNVs conferring increased risk for intellectual disability[49] and schizophrenia[34]. Additionally, the study of rare CNVs revealed that a single CNV might be associated with a wide range of clinical phenotypes. For example, in 2008 a recurrent microdeletion at 15q13.3 was separately associated with schizophrenia[34,50], autism and other neuropsychiatric features[51], and epilepsy and mental retardation[33]. In 2009, this recurrent 15q13.3 microdeletion was tested in a cohort of common epilepsy patients, and found to confer increased risk for the genetic generalized epilepsies (GGE)[31]. This "critical region", or minimum deleted region across all observed patients, is 1.5Mb. This 1.5Mb region contains seven genes, including a plausible epilepsy candidate gene – *CHRNA7*, which encodes a subunit of the nicotinergic acetylcholine receptor. Two additional recurrent CNVs at 15q11.2[52] and 16p13.11[52],[30] have also been associated with the common epilepsies. The microdeletion at 16p13.11 is found in genetic generalized epilepsy and partial epilepsy patients[30].

In summary, there are three recurrent CNVs that increase epilepsy risk: 15q11.2, 15q13.3, and 16p13.11. These microdeletions are especially important for the genetic generalized epilepsies (GGE) and they are also shared risk factors for other neuropsychiatric disorders such as schizophrenia, autism, and intellectual disability. Collectively, these three CNVs account for an estimated 2.9% of patients with genetic (a.k.a. idiopathic) epilepsies[53].

Non-recurrent CNVs are also potential risk factors for all types of epilepsy. No single non-recurrent CNV will account for a large proportion of epilepsy patients. However, these CNVs may include known epilepsy genes or may highlight novel candidate genes (CNVs with different end-points may have a "critical region" that impacts the same novel gene). For example, in a study of ~500 epilepsy patients, two patients harbored microdeletions involving *AUTS2* (previously associated with autism) and one harbored a microdeletion involving *CNTNAP2* (previously associated with autism, Cortical dysplasia-focal epilepsy syndrome, and Pitt-Hopkins like syndrome 1), making these genes relevant to epilepsy given the association of these genes with other neurodevelopmental & neuropsychiatric disorders[53]. Investigations of CNVs in epileptic encephalopathy found that ~4% of patients harbor rare and clearly pathogenic CNVs[54]. More generally, large heterozygous deletions, (>1Mb) are significantly enriched in epilepsy patients, and completely absent from controls when >2Mb in

size[30]. Proving the causality of these rare or singleton CNVs will be difficult; likely requiring very large sample sizes and/or independent evidence for the candidate gene (e.g., association of non-CNV mutation).

Currently, the most well established epilepsy associated CNVs are all deletions; however, this does not to exclude the possibility that pathogenic duplications also exist.

1.3.6.2 Mechanisms of copy number variant pathogenicity

It is clear that phenotypic heterogeneity is associated with risk CNVs and it is also clear that some CNVs are not completely penetrant. The mechanism of pathogenicity for CNVs is not clear and may vary depending on the locus itself (location in the genome) or the individual genome of the patient. In many cases a phenotype may be attributable to a single gene within the CNV; for example a microdeletion of *SCN1A* in a Dravet syndrome patient. For other microdeletions the phenotype may be simply due to haploinsufficency (only one copy of the gene does not result in enough of the gene product) of all the genes within the CNV. Alternately, there may be a deleterious variant present in a gene on the non-deleted homologous chromosome leaving no wild type copies of the gene. Regardless of the mechanism, rare copy number variants play an important role in epilepsy susceptibility.

1.3.7 The role of large structural variants in epilepsy

Large structural variants (>3Mb) can be detected by cytogenetics, which is microscopic analysis of chromosomes in individual cells. There are three main cytogenetic analyses performed in the laboratory: G-banding karyotypes, fluorescence in situ hybridization (FISH) and chromosomal microarrays. Chromosomal microarrays provide the highest resolution of the three techniques and can detect submicroscopic abnormalities that are too small to be detected by conventional karyotyping (i.e., CNVs).

These cytogenetically detectable variants are less frequent in the human genome than CNVs and are often pathogenic. Cytogenetic analysis is particularly helpful for epilepsies occurring with mental retardation or dysmorphic features. Eight syndromes involving seizures are caused by a recurrent chromosomal abnormality, these are referred to as the "chromosomal epilepsies" and include: Down syndrome, Angelman syndrome, Miller–Dieker syndrome, Wolf–Hirschhorn syndrome (4p deletion), Chromosome 1p36 deletion syndrome, 15q inversion-duplication, Ring chromosome 14, and Ring chromosome 20[55]. Additionally, Fragile X patients also frequently experience seizures. This disorder is often discussed amongst the chromosomal epilepsies because early research revealed that these patients frequently had a detectable fragile site on the X chromosome, at which the chromosome was prone to breakage. We now know that this disorder is caused by mutations in the *FMR1* gene, most commonly an expanded

CGG triplet repeat mutation. Thus, despite a commonly detectable abnormality in the X chromosome, fragile X is actually classified as a trinucleotide repeat disorder.

1.3.8 Next-generation sequencing studies in epilepsy

To date, a number of different NGS studies have been conducted with the goal of uncovering new risk factors for epilepsy and epilepsy related syndromes. These are described below.

1.3.8.1 Study design: case-control

A case-control study of 118 genetic generalized epilepsy (previously known as idiopathic generalized epilepsy) cases and 242 controls of European ancestry were exome-sequenced to evaluate the role of rare variants of relatively large effect, that are frequent enough to be present in multiple cases[56]. Specifically, this cohort of 118 GGE cases included 93 juvenile myoclonic epilepsy patients and 25 absence epilepsy patients. Despite restricting the tested variants to SNVs with a MAF<5% that were predicted to disrupt the protein-coding sequence, this exome-sequencing based association testing failed to identify any variants that were significantly associated with GGE. Additionally, a second stage of this study went on to genotype a subset of the SNVs identified by exome-sequencing (n=3,897) in a larger case-control cohort of 878 GGE cases and 1,830 controls; this also failed to identify variants significantly associated with GGE. This work highlights the extreme genetic heterogeneity of epilepsy disorders and also

suggested that gene-based analyses (as opposed to variant-based) and/or more homogeneous phenotypic cohorts will be needed to reveal true risk factors for the genetic generalized epilepsies.

1.3.8.2 Study design: Mendelian families

There have been a handful of other NGS studies in epilepsy; primarily these have investigated specific epilepsy syndromes with familial inheritance patterns. For example, multiple families with autosomal dominant familial focal epilepsy with variable foci (FFEVF) had previously shown linkage to 22q12, but the casual gene had yet to be identified. Therefore, two different research laboratories took the same approach; by: 1) selected families with linkage to this region, 2) performed exome-sequencing on one or more family members 3) identified rare protein-coding variants in this linkage region and 4) tested these variants for cosegregation in the whole family. This resulted in the identification of causal mutations in *DEPC5*[38,39].

Another example is benign familial infantile epilepsy (BFIE), an autosomal-dominant seizure disorder where many families showed linkage to16p11.2-16q12.1 but the causal genetic mutations evaded discovery for many years. Therefore, one research group designed a targeted capture for genes in this linkage peak, ultimately resulting in the identification of *PRRT2* as the causal gene[40]. However, this discovery was actually only due to Sanger sequencing of the *PRRT2* gene because the coverage (number of

short sequencing reads at a given site) was too low to accurately call variants in this gene[40]. This highlights one of the technical pitfalls of NGS: coverage must be relatively high (typically >30x on average across the genome) to accurately detect variants and this can be difficult in certain regions of the genome (e.g., GC-rich).

1.3.8.3 Study design: trios (focused on *de novo* mutation detection)

The critical contribution of *de novo* mutations to neurodevelopmental disease risk has recently been elucidated[57-59]. Recently, the role of *de novo* mutations in epileptic encephalopathies has investigated. As a member of the Epi4K Consortium[60], one such project will be covered in Chapter 4 [42]. Additionally, a number of other studies have used targeted capture to sequence only a subset of relevant genes. One such study sequenced nine known and 46 candidate genes for epileptic encephalopathy in 500 cases[41]; resulting in the association of epileptic encephalopathy with *de novo* mutations in two novel genes: Chromodomain Helicase DNA Binding Protein 2 *(CHD2)* and Synaptic Ras GTPase Activating Protein 1 *(SYNGAP1)*[41]. A second study sequenced 35 known or potential candidate genes in 53 epileptic encephalopathy patients[61]; resulting in the identification of a number of causal *de novo* mutations in previously known genes.

When looking at epilepsy more generally, there has still been a strong focus in the community on the identification of *de novo* mutations. This has resulted in an

increasing number of candidate genes. In fact, in a Pubmed search (June 20th, 2013) for the term 'De novo mutations + seizure' returned 244 primary papers highlighting a gene or structural variant linked to clinical manifestations of seizure. Within these papers, 65 unique genes have been identified, as well as 91 structural variants. Several genes were reported in multiple papers, including *SCN1A* (35), *CDKL5* (11), *PCDH19* (9), and *KCNQ2* (7). Further classification of these genes revealed that 11 are known epilepsy genes[27], 11 are associated with epilepsy in the Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/ac/index.php), 16 are included on an epilepsy NGS gene panel[43], eight are listed with an association to a known "condition with seizures" on Genetics Home Reference (http://ghr.nlm.nih.gov), two genes harbor *de novo* mutations in patients with Autism Spectrum Disorder[58,59,62,63], one is a mouse seizure susceptibility gene[64], four are known ion channel genes[65], and 12 do not fall into any of these categories.

While many of these *de novo* mutations are likely disease causing, each *de novo* mutation should be considered on a case-by-case basis. Generally, a coding mutation that is predicted to be deleterious in a previously established gene is considered causal. For newly implicated genes, additional observations of mutations in other phenotypically similar patients are necessary in order to prove causality; a good framework for such analyses has recently been established[42]. As sequencing continues

to drop in cost and as sample sizes grow, the discovery of additional *de novo* mutations will help to widen our understanding of the genetic basis of epilepsy and other related seizure conditions.

1.3.9 Epilepsy pharmacogenomics

Genetic variation can not only confer disease risk or protection, but also influence how a patient responds to medications. The study of the role of genetics in pharmacologic response, or pharmacogenetics, is particularly relevant in epilepsy disorders since nearly all epilepsy patients undergo drug therapy for some period of time, and an estimated 30% of patients are drug resistant and never achieve seizure freedom despite treatment with all available antiepileptic medications.

Several genetically based hypotheses exist for why some patients fail medications that can control seizures in other patients with the same diagnosis. One hypothesis, the pharmacokinetic hypothesis, is that genetic variation causes differences in the absorption, metabolism, or distribution of antiepileptic drugs in the brain[66]. Under this model, a drug-responsive patient receiving a medication will get adequate concentrations of medication to the brain, whereas a drug-resistant patient will be unable to achieve sufficient medication levels in the brain to achieve seizure control. To date, the only consistently replicated evidence supporting this hypothesis is the association of variants in genes encoding a drug-metabolizing enzyme, CYP2C9, with

phenytoin (a common antiepileptic drug) metabolism[67-69]. Given the relatively minor consequences of the variation on the antiseizure response of phenytoin, the clinical utility of using *CYP2C9* genotype to predict phenytoin dosage requirements is unclear. Despite this uncertain clinical relevance, dosing adjustments have been published recommending a 25% reduction in maintenance dose for the CYP2C9 *1/*2 and *1/*3 genotypes, a 50% reduction in maintenance dose for the CYP2C9 *2/*2, *2/*3, and *3/*3 genotypes, and cautious monitoring for adverse drug reactions associated with phenytoin toxicity (sedation, ataxia, nystagmus, dysarthria)[70]. An alternative hypothesis is the pharmacodynamic hypothesis, where the gene encoding a drug target is mutated, and this change prevents the medication from being able to effectively modulate the intended pathway[71]. There are no consistent reports of mutations in genes encoding drug targets or their modulators associating with response to antiepileptic dosing. It should be noted, however, that pharmacogenetic studies of antiepileptic medications performed to date have primarily evaluated the role of common variants (MAF >5% in the population) and are often statistically underpowered to robustly detect genetic associations. Larger studies considering less common variation are needed to better explore the aforementioned hypotheses.

In addition to there being little evidence to support either the pharmacokinetic or pharmacodynamic genetic hypotheses, they also fail to explain why a large number of

drug-resistant patients fail to respond to multiple medications that are substrates for many different transporters at the blood-brain-barrier and have differing modes of pharmacologic action. That is, under both of these hypotheses, many universally drug-resistant patients would have to have multiple mutations in genes encoding drug transporters governing drug disruption, metabolic pathways mediating metabolism, and/or pharmacologic targets. While it is possible that some patients may have acquired diffuse non-genetic changes due to seizures that are responsible for multi-drug resistance, some patients present at the onset with drug-resistant seizures, and despite aggressive early interventions, fail to respond. Recently, it has been proposed that drug-resistant epilepsy may reflect an intrinsically more severe form of epilepsy and that seizure control is therefore more difficult from the onset[72]. One interesting additional possibility stemming from this hypothesis is that certain forms of epilepsy are not necessarily intrinsically more severe, but rather they arise from pathophysiologic changes that are not correctable with current pharmacologic agents. Therefore, what appears to be more severe epilepsy, may in fact be a specific subtype of epilepsy with a currently unknown biologic etiology. Additional research is needed to reveal the neurobiological and genetic aspects of drug-resistance, particularly as it pertains to the interplay of drug-resistance and underlying pathophysiology.

Genetic variation can also dictate if a patient will experience severe adverse drug reaction to a medication. For antiepileptics, two clear genetic associations of HLA (human leukocyte antigen) alleles with cutaneous hypersensitivity reactions have been identified. First, in 2004 Chung *et al.* first reported a strong association between the presence of the HLA-B*1502 allele and carbamazepine-induced Stevens-Johnson syndrome, a severe hypersensitivity reaction, in patients of Han Chinese descent. In this study, the HLA-B*1502 allele was present in 100% (44/44) of patients with carbamazepine-induced Stevens-Johnson syndrome, while only 3% (3/101) of patients on carbamazepine with the HLA-B*1502 allele had no reaction[73], indicating that the presence of this allele is highly predictive of this severe reaction to carbamazepine. These results were later replicated in the Han Chinese population, in other Asian populations, and also expanded to include toxic epidermal necrolysis, another severe cutaneous hypersensitivity reaction[74]. Based on these findings and the frequency of the HLA-B*1502 alleles in Asian populations, the Food and Drug Administration currently recommends that all high risk populations, including individuals of Han Chinese descent, and individuals from Vietnam, Cambodia, the Reunion Islands, Thailand, India, Malaysia, and Hong Kong, be screened for the presence of the HLA-B*1502 allele prior to initiating carbamazepine drug therapy. Limited evidence suggests that the HLA-B*1502 allele may also increase the risk of phenytoin-induced severe

hypersensitivity reactions[75], which has led to the inclusion of a statement of caution in the drug label for phenytoin.

In addition to the risk of hypersensitivity reactions associated with the HLA-B*1502 allele, two recent genome-wide association studies in patients of European and Japanese ancestry showed an association of an allele of the gene encoding the human leukocyte antigen A (HLA-A*3101) was associated with a range of carbamazepine-induced hypersensitivity reactions including: maculopapular exanthema, hypersensitivity syndrome, and the more severe Stevens-Johnson syndrome and toxic epidermal necrolysis[76,77]. This association was not as strong as that seen for HLA-B*1502, with only moderately increased risk for hypersensitivity reactions[76,77]. Unlike HLA-B*1502, HLA-A*3101 does not increase the risk of hypersensitivity to other aromatic antiepileptic medications, including lamotrigine and phenytoin[78]. There are not currently recommendations on genotyping HLA-A prior to carbamazepine dosing.

1.3.10 Genetic testing of epilepsies in the clinic

There are several varieties of genetic testing used in a clinical setting; including: biochemical assays, cytogenetics, Sanger sequencing (typically used for single gene testing), and next-generation sequencing (used for gene panel sequencing or whole-exome sequencing). Genetic testing can sometimes determine a genetic diagnosis, which in turn, helps guide patient care and counseling about prognosis and reproductive

choices for the parents. In addition to diagnostic testing, predictive testing may be valuable for patients with a family history of epilepsy. In the United States, all genetic testing that will be used for medical management or intervention for a patient must be performed in a Clinical Laboratory Improvement Amendments (CLIA) certified laboratory. Access to a genetic counselor should be provided to the patient and their family for a discussion of the results and their implications.

It is critical to be aware of the current speed of genetic discovery because this dictates the growth in the field of genetic testing, with new information emerging practically every day. In epilepsy genetic testing, there are guidelines established for the clinical utility of testing the most commonly mutated genes[27,55]. For example, screening of *SCN1A* in Dravet syndrome or *CDKL5* in infantile spasms are useful screens because mutations (typically *de novo*) explain 70-80% and 10-17% of cases, respectively[27]. In patients experiencing seizures with cortical malformation, one of the most common diagnoses is periventricular nodular heterotopia (PNH). The most common cause of familial PNH and ~25% of the nonfamilial forms are caused by mutations in the *FLNA* gene; again making this a logical first diagnostic screen[55]. For a small proportion of genes, knowing the genetic etiology can guide treatment as is the case for *SLC2A1* mutation positive patients ("GLUT1 deficiency syndrome") where the ketogenic diet is the gold standard of treatment[79].

Chromosomal microarray analysis and karyotyping should be used to detect structural abnormalities; this is particularly critical for patients with seizures who exhibit mild or moderate intellectual disability and/or dysmorphic features. In this case, fragile X testing and biochemical testing of amino acid levels may also be useful[55]. CNV screening appears to be critical in two main groups: epileptic encephalopathy (CNVs explain ~4% of cases)[54] and GGE with intellectual disability (CNVs explain ~10% of cases)[80].

In patients where the first genetic screen is less obvious, or if screening of multiple genes has failed to identify the disease-causing variant, ordering a comprehensive panel of genes is a very practical approach. For many diseases, including epilepsy, gene panels are offered for genetic testing; gene panels target a set of genes that are associated with a given disorder. Sequencing of these genes is usually achieved by next-generation sequencing of the coding exons and the flanking intronic boundaries. The set of genes included may vary by the company performing the CLIA certified sequencing. For example, Transgenomic® offers a Comprehensive Epilepsy Evaluation NGS Panel which includes 377 genes that have been reported to cause epilepsy or seizure related disorders. Courtagen® offers a similar panel of 327 genes and recently announced they will expand this list to 489 genes associated with seizure disorders and offer a panel for "Infancy and Childhood Epilepsy" targeting 70 genes (although these

new panels are not listed on their website as of 2/26/14). Current lists of all available genetic tests and testing facilities are available online and updated regularly (http://www.genetests.org and http://www.orpha.net).

From a clinical perspective, it is more practical to order a phenotypically relevant gene panel than to order exome sequencing. One reason for this is that variants identified in genes selected for a gene panel will be more interpretable based on prior knowledge; in contrast, interpretation of whole exome sequence data may generate a long list of variants of "unknown clinical significance". Additionally, this avoids issues associated with "incidental genetic findings" which may arise when interrogating the exome. However, if the patient has a very unique phenotype and good medical insurance, exome-sequencing may be a more desirable approach than a gene panel. Finally, it is usually helpful to also have parental DNA available for testing to determine the inheritance (i.e., *de novo* status) and thus likely pathogenicity of any identified variants.

1.3.11 Epilepsy genetics summary

The scientific community has made important strides toward understanding the genetic etiology of the epilepsies. First, only a few rare types of epilepsy are caused by mutations in single genes where these mutations segregate in families according to Mendel's laws. The majority of genes in this category have likely been identified

already, at least if these families have been ascertained already. Any remaining cases of "low hanging fruit", such as families with linkage peaks (like *DEPC5*[38,39]), will likely be solved in the near future with NGS. Collectively, mutations in these genes explain an estimated 1% of epilepsy cases. Second, the findings from these families have informed the biology of epileptogenesis and shown that diverse classes of genes can lead to sporadic epilepsy and syndromic forms of epilepsy. The genotype-phenotype correlations observed in these genes also prove the locus heterogeneity and variable expressivity that confound traditional genetic approaches. Third, despite the lack of clear associations between common variants and epilepsies, they have informed our understanding of the genetic architecture and focused our efforts towards rare variants. Fourth, CNVs also confer increased risk for epilepsy, although the mechanism of pathogenicity is not clear for many of these CNVs. Fifth, large-scale exome-sequencing of GGE cases failed to identify any variants of large effect. This suggests association tests at the gene-level are more promising and also highlights the high genetic heterogeneity that likely exists in complex epilepsies. While trio-based NGS studies have been particularly successful in the epileptic encephalopathies, we are still only explaining an estimated ~12% of *SCN1A* mutation negative patients. Based on the large-scale sequencing efforts in IS and LGS, it has been estimated that ~90 genes confer risk[42], again highlighting the extreme genetic heterogeneity for the epilepsies.

2. Deficiency of Asparagine Synthetase Causes Congenital Microcephaly and a Progressive Form of Encephalopathy[1]

2.1 Introduction

Intellectual disability (ID) affects 2%–3% of the general population and is characterized by a broad range of cognitive deficits. It is usually subdivided into syndromic and nonsyndromic forms, depending on whether additional abnormalities are found. Syndromic ID is often accompanied by microcephaly, defined by a head circumference more than two standard deviations (SD) below the age- and sex-adjusted mean. The incidence of microcephaly, as reported in birth defect registries world-wide, varies from 1 to 150 per 100,000 depending upon the range of SD used to define microcephaly and the ethnic population. For example, microcephaly is more prevalent in populations with a high degree of consanguinity [81]. Causes of congenital microcephaly include metabolic disorders, chromosomal anomalies, and intrauterine infections. However, with the exception of autosomal recessive primary microcephaly (MCPH), the genetic etiology of most congenital microcephaly cases is unknown.

We ascertained four families with a distinct form of severe encephalopathy

[1] This chapter is part of a published work [15].

associated with congenital microcephaly and progressive brain atrophy. Two families were from the same ethnic group, whereas the other two families were independently recognized as presenting with an identical syndrome. Both pairs of families were analyzed independently by exome sequencing. This chapter summarizes the clinical features of the affected children and shows that the observed phenotype in all four families can be explained by autosomal recessive deficiency of asparagine synthetase (ASNS).

2.2 Materials and Methods

2.2.1 Recruitment of Subjects and Collection of Samples

Families A and B were recruited at Sheba and Wolfson Medical Centers in Israel, family C at The Hospital for Sick Children in Toronto (Canada), and family D at Sainte-Justine Hospital in Montreal (Canada). Blood samples were obtained from most affected individuals, their unaffected siblings, and their parents. The relevant Institutional Review Boards approved the studies and appropriate family members gave written consent.

2.2.2 Sequencing and Variant Identification

Exome Sequencing in Families A and B

The Illumina Genome Analyzer IIx platform (Illumina) was used to perform exome sequencing of three microcephaly patients (Figure 5, family A: II.1 and family B:

II.2 and II.4) at the Center for Human Genome Variation (CHGV) at Duke University, Durham, NC. Prior to sequencing, target regions were captured using the SureSelect Human All Exon technology (Aligent Technologies). This technology captures consensus coding sequence exonic regions and flanking intronic regions totaling ~38 Mb of genomic DNA. The resulting short-sequence reads were aligned to the reference genome (NCBI human genome assembly build 36; Ensembl core database release 50_361 [82]) using the Burrows-Wheeler Alignment (BWA) tool [83]. After accounting for PCR duplicates (removed using the Picard software: http://picard.sourceforge.net) and reads that did not align to captured regions of the reference genome, the average coverage for these three samples was ~71x and each sample had >95% of the bases covered. A base within the 37.8 Mb captured region was defined as covered if ≥5 short reads spanned this nucleotide (Table 3).

Table 3. Sequencing coverage for samples from families A-D.

A base within the 37.8Mb (families A and B) or the 52 Mb targeted exonic regions (families C and D) was defined as covered if ≥5 short-reads spanned this nucleotide. After accounting for PCR duplicates and reads that did not align to captured regions of the reference genome, the average read depth for each sample was greater than 63x and greater than 95% of targeted bases were covered in all eight samples.

Individual	Total covered bases (MB)	% covered bases	Average read depth
Family A (II-1)	36.32	96.1	79.92
Family B (II-2)	36.30	96.0	63.93
Family B (II-4)	36.21	95.8	69.59
Family C (II-3)	51.28	95.0	80
Family D (II-1)	51.64	97.5	144
Family D (II-2)	51.37	95.8	126
Family D (II.4)	51.53	96.3	128
Family D (II.5)	51.67	97.8	89

Genetic differences between each patient genome and the reference genome were identified using the SAMtools variant calling program [84], which identifies both single-nucleotide variants (SNVs) and small indels. We then used the Sequence Variant Analyzer software (SVA) [85] to annotate all identified variants. SVA was also used to apply quality control filters to the variants identified by SAMtools. High-quality SNVs were obtained using the following criteria: consensus score ≥20, SNP quality score ≥20, and reads supporting SNP ≥3. High-quality indels were obtained using the following criteria: consensus score ≥20, indel quality score ≥50, ratio of (reads supporting variant/reads supporting reference): 0.2–5.0, and reads supporting indel ≥3.

Exome Sequencing in Families C and D

The exomes of one of the individuals of family C (Figure 5: II.3) and four individuals of family D (Figure 5: II.1, II.2, II.4, and II.5) were captured using the Agilent SureSelect all exon kit V3 (approximately 51.9 Mb of target sequences) and then sequenced in pair ends (2 x 100 bp) on the Illumina HiSeq2000 (v3 chemistry; 3 exomes/lane format) at the McGill University Genome Quebec Innovation Center (Montreal, Canada). The sequences were aligned and the variants were called using GATK [86]. After removal of duplicate reads, using Picard, we obtained an average coverage of >80x per target base, with 95% of the target bases being covered at ≥10x (Table 3). Only variants that meet all the following criteria were considered: base coverage ≥8x, reads supporting the variant ≥3, and ratio of reads supporting variant/reads supporting reference ≥20%. Variants were then annotated using Annovar [87].

2.2.3 Variant validation and control genotyping

Genotyping p. F362V

Genotyping of the p.F362V variant in 1,160 controls was performed in the Center for Human Genome Variation at Duke University (Durham, NC). This was done using a custom TaqMan genotyping assay (Applied Biosystems): forward: 5′–CCT GCG TAA

GTT CAT CTG ATC CTT-3′; reverse: 5′–GTA TAT TCG GAA GAA CAC AGA TAG CGT-3′; probe: 5′-TCC AGA GA[A/C] GAT CAC C-3′.

Genotyping of the p.F362V variant in 80 Iranian Jewish controls and the non-exome-sequenced family members (Figure 5: family A: I.1, I.2, II.2, II.3, and II.4 and family B: I.1, I.2, II.1) was performed at the Gertner Institute of Human Genetics, Sheba Medical Center, Israel. Sanger Sequencing (Figure 7) or restriction digest with the restriction enzyme Alw26I (data not shown) were used to perform this genotyping. Both methods used the following custom primer sequences: forward: 5′-CTT TCA ATT ATT TCC AAA AAT CAA ATC-3′ and reverse: 5′–CAC TGT CAT ACT GAA AGA TGA TAG AAA-3′. These primers resulted in a 286 base pair amplicon that targeted the nucleotide of interest.

The p.F362V variant, found in families A and B, was validated in these three samples using all three methods: TaqMan genotyping, Sanger sequencing, and restriction digestion.

Genotyping p.R550C and p.A6E

Sanger sequencing of PCR-amplified products was used to genotype p.R550C and p.A6E variants (Figure 7). The following custom primers were used for p.A6E: forward: 5′ – GCC GGT TGA ATG TAG AGG TC-3′ and reverse: 5′ – CCA AAG CAG CAG TTG GTG TA-3′. The following custom primers were used for p.R550C: forward: 5′

- GCC ATT TTA AGC CAT TTT GC-3′ and reverse: 5′ – TTT CCC TTT TCC TAG CTT ACC C-3′. The mutations p.R550C and p.A6E were genotyped in 300 French Canadian healthy controls. In addition, p.R550C was genotyped in 225 Bangladeshi healthy controls.

2.2.4 Haplotype prediction

To determine if the p.F362V *ASNS* variant was always found on the same haplotype, all high confidence SNVs (coverage ≥10, ≥5 reads supporting the variant, SNV quality ≥30, SNV consensus ≥30) on chromosome 7 were collected from each of the three exome-sequenced patients (Families A and B). Next, the regions on either side of the *ASNS* variant were examined to find the boundaries of the homozygous stretches containing the *ASNS* variant in each individual. Finally, the SNVs in this region were compared across samples to obtain the largest possible haplotype that was shared by all three affected patients. This estimated haplotype is approximately 1.2 Mb (build 36, Chr7: 97322654-98488035) and is tagged by 16 SNVs (Table 11).

These 16 SNVs were also assessed in the 261 sequenced controls. The genotype for each of these SNVs was input into the fastPHASE program[88] in order to infer the phase of these variants in each control sample. Only 17 controls had all 15 SNVs (the 16th variant is the *ASNS* variant at the outer boundary, which was not found in any controls), and the majority of these variants were found in heterozygous form. Of the 522 possible

haplotypes, fastPHASE determined that 2 haplotypes were identical to the *ASNS* haplotype (excluding the *ASNS* variant itself) when the individual haplotype error was minimized, and 4 haplotypes were identical to the *ASNS* haplotype (again, excluding the *ASNS* variant itself) when switching error was minimized. All 4 of these control samples are self-identified as Caucasian.

2.2.5 Homozygosity mapping

To identify shared regions of homozygosity between the families A and B, we used the exome-sequence data to detect homozygous regions present in these three cases. We used high-quality single nucleotide variants (coverage ≥10, ≥5 reads supporting the variant, SNV quality ≥30, SNV consensus ≥30) identified in exome-sequencing to perform homozgyosity mapping. The SNVs in all three of the cases were analyzed using the PLINK "homozygosity mapping" tool [89]. This algorithm considers sliding windows across the genome and screens for homozygous variants in those regions. To account for the nonrandom spacing of SNVs in the exome data and to detect such regions with high sensitivity, we defined a "window" as 10 SNPs within a 1,000 Kb window, allowed one heterozygous SNV per window, and defined a homozygous segment as a region with 10 SNVs in a 1 Kb window.

Homozygosity mapping was also done in family C at the McGill University Genome Quebec Innovation Center (Montreal, Canada). Briefly, genomic DNA (~2.5µg)

from the blood of the two affected siblings (C:II.3 and II.4) was used to perform whole-genome genotyping using the Illumina 660 Quad Chip (Infinium HD). PLINK (v.1.06) [89] was used to identify shared regions of homozygosity with a size of >1Mb.

2.2.6 Clone Preparations

Full-length cDNA encoding human ASNS was amplified from first-strand cDNA derived from the HEK293 human kidney cell line with an RNeasy plus mini kit (QIAGEN), High Capacity cDNA Reverse Transcription Kit (Applied Biosystems), Phusion HF DNA polymerase (Finnzymes), and a specific primer set (5′ – CTC GAG ATG TGT GGC ATT TGG GCG CT - 3′ and 5′ – CTC GAG CCT AAG CTT TGA CAG CTG ACT - 3′). The cDNA was subcloned into the pCR-Blunt II-TOPO vector (Invitrogen-Life Technologies) and subjected to sequence analysis (pCR-Blunt II-ASNS-WT). Using pCR-Blunt II-ASNS-WT, A6E, F362V, and R550C of ASNS were made by PCR-mediated site-directed mutagenesis using Phusion HF DNA polymerase and a specific primer set (A6E: 5′- GCT GTT TGG CAG TGA TGA TTG -3′ and 5′ –TCC CAA ATG CCA CAC ATC TC -3′ ; F362V: 5′ – GTC TCT GGA GAA GGA TCA GA-3′ and 5′ – GAT CAC CAC GCT ATC TGT GT-3′ ; R550C: 5′ – GCA CGC TGA CCC ACT AC -3′ and 5′ – AGG CAG AAG GGT CAG TGC-3′), which were phosphorylated by T4 polynucleotide kinase (New England BioLabs). The amplicons were self-ligated using T4 DNA ligase (Promega) and subjected to sequence analysis (pCR-Blunt II-ASNS-A6E,

pCR-Blunt II-ASNS-F362V, and pCR-Blunt II-ASNS-R550C). ASNS human cDNA containing each allele was subcloned into the pcDNA3.1(+) vector (Invitrogen-Life Technologies) using the KpnI and XbaI sites from pCR-Blunt II-ASNS-WT, pCR-Blunt II-ASNS-A6E, pCR-Blunt II-ASNS-F362V, or pCR-Blunt II-ASNS-R550C and subjected to sequence analysis (pcDNA3.1(+)-ASNS-WT, pcDNA3.1(+)-ASNS-A6E, pcDNA3.1(+)-ASNS- F362V, or pcDNA3.1(+)-ASNS-R550C; Figure 3).

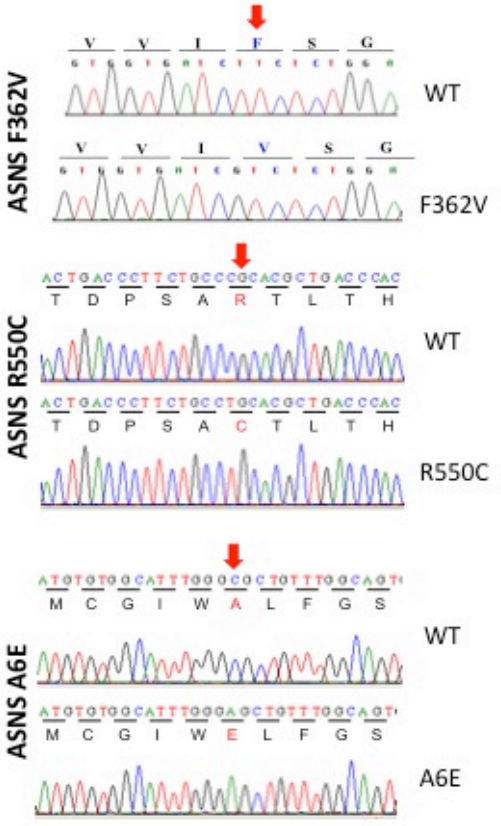


Figure 3. Sanger sequencing of in vitro ASNS alleles.
After PCR-directed mutagenesis, cDNA encoding human *ASNS* containing wild type or mutant alleles was subcloned into pcDNA3.1(+) vector and subjected to sequence analysis.

Using pcDNA3.1(+)-ASNS-WT, pcDNA3.1(+)-ASNS-A6E, pcDNA3.1(+)- ASNS-F362V, or pcDNA3.1(+)-ASNS-R550C, FLAG-tagged-modified ASNS was made by two-step PCR-mediated site-directed mutagenesis using Phusion HF DNA polymerase and specific primer sets (first step: 5′- GAC AAG TAG GCT CGA GAA GGG -3′ and 5′ – GTA GTC AGC TTT GAC AGC TGA C -3′ ; second step: 5′- GAC GAT GAC AAG TAG GCT CGA GAA GGG -3′ and 5′- GTC CTT GTA GTC AGC TTT GAC AG -3′), which were phosphorylated by T4 polynucleotide kinase, the amplicons were self-ligated using T4 DNA ligase and subjected to sequence analysis (pcDNA3.1(+)-ASNS-FLAG-WT, pcDNA3.1(+)-ASNS-FLAG-A6E, pcDNA3.1(+)-ASNS-FLAG-F362V, or pcDNA3.1(+)-ASNS-FLAG-R550C). cDNAs encoding FLAG-tagged human ASNS were subcloned into pcDNA3.1(+) vector again, using the KpnI and XbaI sites and subjected to sequence analysis (Figure 4).

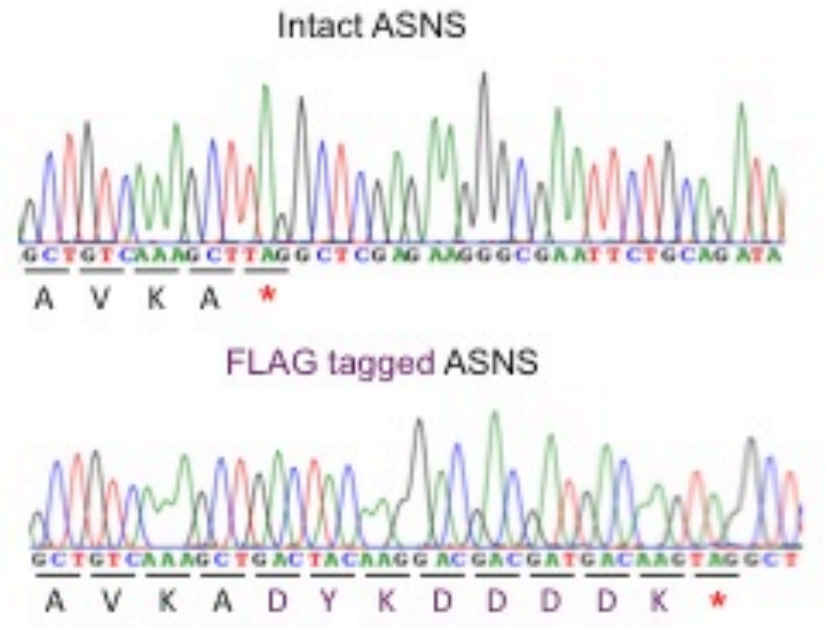


Figure 4. Sanger sequencing of FLAG-tagged ASNS.
A FLAG-tag was introduced into the sequence encoding the C-terminus of ASNS for the wild type and the three mutant forms (A6E, R550C and F362V) and subjected to sequence analysis.

2.2.7 Cell Culture: RT-PCR

Empty pcDNA3.1 (+) vector, pcDNA3.1(+)-ASNS wild-type, or pcDNA3.1(+)-ASNS mutant (p.F362V, p.R550C, or p.A6E) were transfected into the monkey COS-7 kidney cell line or human HEK293 kidney cells by lipofection using Lipofectamine 2000 (Invitrogen-Life Technologies). Total RNA was extracted from transfectants using an RNeasy plus mini kit, and first-strand full-length cDNA encoding human ASNS was synthesized using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). RT-PCR to detect ASNS mRNA expression was performed in 25 cycles at

96˚C for 30 s, 60˚C for 30 s, and 72˚C for 30 s using AmpliTaq Gold DNA polymerase (Applied Biosystems) and a specific primer set (5′- TGC ACG CCC TCT ATG ACA AT -3′ and 5′- CAC CTT TCT AGC AGC CAG TA -3′) (Figure 8(A)).

2.2.8 Cell Culture: Western Blotting

Forty-eight hours after transfection, the cells were lysed with RIPA buffer (Sigma-Aldrich) with Protease inhibitor cocktail (Sigma-Aldrich), and the lysates were subjected to SDS-PAGE gel and transferred to a polyvinylidene difluoride membrane (Millipore). The membranes were incubated with anti-FLAG M2 monoclonal antibody (Sigma-Aldrich) or anti-actin antibody (Santa Cruz Biotechnology). Proteins were visualized with the ECL plus western blotting detection system (GE Healthcare). For Leupeptin treatment, 24 hr posttransfection, the cells were incubated with 100 mM Leupeptin (Sigma Aldrich). After 8 hr incubation with Leupeptin, the cells were lysed, and FLAG-tagged ASNS were detected as above.

2.2.9 Protein Conservation

Species and ASNS proteins were from gi P08243 (Human, Homo sapiens), ENSMUSP00000031766 (mouse, Mus musculus), ENSGALP00000015846, (chicken, Gallus gallus), ENSACAP00000012780 (lizard, Anolis carolinensis), ENSXETP00000054608, (frog, Xenopus tropicalis), ENSTRUP00000013503 (fish, fugu, Takifugu rubripe), FBpp0089009 (fruit fly, Drosophila melanogaster), NP_741864 (worm,

nematode, Caenorhabditis elegans), YGR124W (yeast, Saccharomyces cerevisia), and YP_003233213.1 (bacterium, Escherichia coli) (Figure 11). Sequence alignment was performed using ClustalW [90] and alignment editing with the BioEdit software (http://www. mbio.ncsu.edu/bioedit/bioedit.html).

The pdb structure was made using Discovery Studio program (http:// accelrys.com/products/discovery-studio) (Figure 11).

2.2.10 Mouse Analyses

Genetically modified *Asns* mice were obtained from the Eucomm consortium. Mice were maintained by breeding to C57BL/6NTac. Mice were genotyped using the following custom primer sequences: forward: 5′- TTT TGG TTT GTG TTT CTT CCT G -3′ and reverse: 5′- TCA GGA ACG TGA GTG AGT GAG T -3′. Histology at P0 was performed by cryopreservation of tissue, cryosectioning, and hematoxylin and eosin staining. Histology in adult brains was performed by fixation of tissue using formalin perfusion. Tissue was sent to http://www.histoserv.com for paraffin embedding, sectioning, and staining. Analysis of area and thickness was performed by quantifying measurements using ImageJ. The p-values for structural measurements were obtained using an unpaired t test and calculations were done using R.

cDNA

Mouse cerebral hemispheres were carefully dissected. Total RNA was extracted from brain tissue using an RNeasy plus mini kit, and first-strand full-length cDNA encoding human *ASNS* was synthesized using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems).

Quantitative Real-Time RT-PCR

Quantitative real-time PCR was done using an Asns gene expression assay, with FAM reporter, spanning exons 7–8 (Mm00803785_m1; Life Technologies) and a Gapdh gene expression assay with VIC reporter (Mm99999915_g1; Life Technologies). Samples were run in triplicate and the standard curve was made using cDNA from a non-test wild-type sample. Twelve mice between 3 and 4 months of age were used for qPCR. Four mice of each genotype were used ($Asns^{+/+}$, $Asns^{+/-}$, and $Asns^{-/-}$). One-way ANOVA was used to assess expression differences between the three genotypes ($p < 0.00001$). A post hoc two-tailed t test was then used to assess genotypic differences in expression ($P_{WT\text{-}Asns^{+/-}} = 0.00001$, $*P_{WT\text{-}Asns^{-/-}} < 0.00001$, $*P_{Asns^{+/-}\text{-}Asns^{-/-}} = 0.00083$).

Semi-quantitative RT-PCR

RT-PCR to detect Asns mRNA expression was performed in 35 cycles at 96°C for 30 s, 58°C for 30 s, and 72°C for 90 s using AmpliTaq Gold DNA polymerase (Applied Biosystems) and a specific primer set (5′- CAG TGT CTG AGT GCG ATG AAG A -3′ and

5′- GCG TTC AAA GAT CTG ACG GTA G -3′). RT-PCR to detect Gapdh mRNA expression was performed in 25 cycles at 96°C for 30 s, 57°C for 30 s, and 72°C for 45 s using AmpliTaq Gold DNA polymerase (Applied Biosystems) and a specific primer set (5′- ACC ACA GTC CAT GCC ATC AC - 3′ and 5′- CAC CAC CCT GTT GCT GTA GCC -3′).

Western Blotting

Two different antibodies were tried for detection of mouse Asns: anti-human-ASNS, which recognizes amino acid residues 506–520 of ASNS (Sigma- Aldrich), and anti-Asns, with species reactivity in mouse, rat, and human, which recognizes amino acid residues at the C terminus (Abcam). Both were nonspecific (data not shown).

Mouse behavioral testing

Adult male and female B6NTac (Taconic Labs, Hudson, NY) served as controls and were tested with the heterozygous and homozygous mutant mice (B6NTac;B6N-Asns$^{tm1a(EUCOMM)Wtsi}$/H; European Mutant Mouse Archive, Munich, Germany). Adult CONT (n=8), Asns$^{+/-}$ (n=11), and Asns$^{-/-}$ (n=6) mice were housed under controlled temperature and humidity conditions under a 14:10 h light:dark cycle (light onset 0800 h) with food and water provided ad *libitum*. All behavioral assessments were conducted during the light cycle between 1000 and 1700 h. Mice were tested (CONT (n=8), Asns$^{+/-}$ (n=11), and Asns$^{-/-}$ (n=6)) in following order with 3-5 days interposed between tests:

light-dark emergence, open field, rotorod, and novel object recognition as described [91]. All procedures were approved by the Duke University Institutional Animal Care and Use Committee and were in conducted accordance with NIH guidelines for the care and use of laboratory animals.

All behavioral data are presented as means and SEM and were analyzed with SPSS 20 (IBM North America, New York, New York) using ANOVA (open field, light-dark emergence test), repeated measures ANOVA (RMANOVA; rotarod, novel object recognition test), and the Kruskall-Wallis test for rank-ordered data (percent increases in performance during rotorod testing). Bonferroni corrected pair-wise comparisons were used for *aposteriori* comparisons where a $p<0.05$ was considered significant.

The mice were tested in the light-dark emergence test for 5 min and were monitored with Med-PC IV software (Med-Associates, St. Albans, VT); the open field data were collected over 30 min with Versamax software (Accuscan Instruments, Columbus, OH); the accelerating rotorod data were collected over 5 min with Med-Associates rotorod software; and the novel object recognition testing was conducted in Plexiglas arenas with mouse-friendly objects over 10 min and analyzed with TopScan software (Clever Sys Inc., Reston, VA). Preference scores in the novel object recognition test were calculated by subtracting the total time spent with the novel object from the total time spent with the familiar object during each test, and divided by the total time

spent with both objects. Positive scores indicated a preference for the novel object, negative scores a preference for the familiar object, and scores approaching zero indicated no preference for either object.

Video EEG Recordings of Mice

Two adult Asns homozygous mice and one age-matched WT mouse were anesthetized by intraperitoneal injection of Nembutal (60 mg/kg). Under stereotaxic guidance, four monopolar electrodes were implanted into the subdural space over the left and right parietal cortex and occipital cortex for chronic EEG recording. After a seven day postoperative recovery, EEG activity was recorded with the mouse moving freely in a cage for 80 hr; animal behavior was recorded simultaneously with a digital video camera. An individual blinded to mouse genotype analyzed the EEG and behavioral activities.

2.3 Results

2.3.1 Identification and validation of *ASNS* mutations

We identified a total of nine children from four families with a severe form of intellectual disability (Figure 5; Table 4; Appendix A). These children were born with a small head circumference and showed progressive microcephaly. Although congenital microcephaly is a consistent feature of this syndrome, the patients do not fit the definition of primary microcephaly (MCPH) (Appendix A). Their clinical course was

characterized by profound developmental delay and, in a majority of cases, early-onset intractable seizures (Figure 5; Table 4; Figure 6). Clinical examination revealed axial hypotonia with severe appendicular spasticity in all cases. All affected siblings of family C also showed excessive startle reflex, mimicking hyperekplexia. In addition, several affected individuals from families C and D had episodes of hypothermia. Brain MRI first performed in early infancy showed decreased cerebral volume and size of pons, presumably caused by hypodevelopment and/or atrophy, as well as delayed myelination (Figure 6; Appendix A). Some patients also showed gyral simplification (Appendix A). The affected children from two families (C and D) died during the first year of life because of pulmonary aspiration secondary to severe neurological dysfunction, whereas the affected individuals from the other families survived into their third decade.

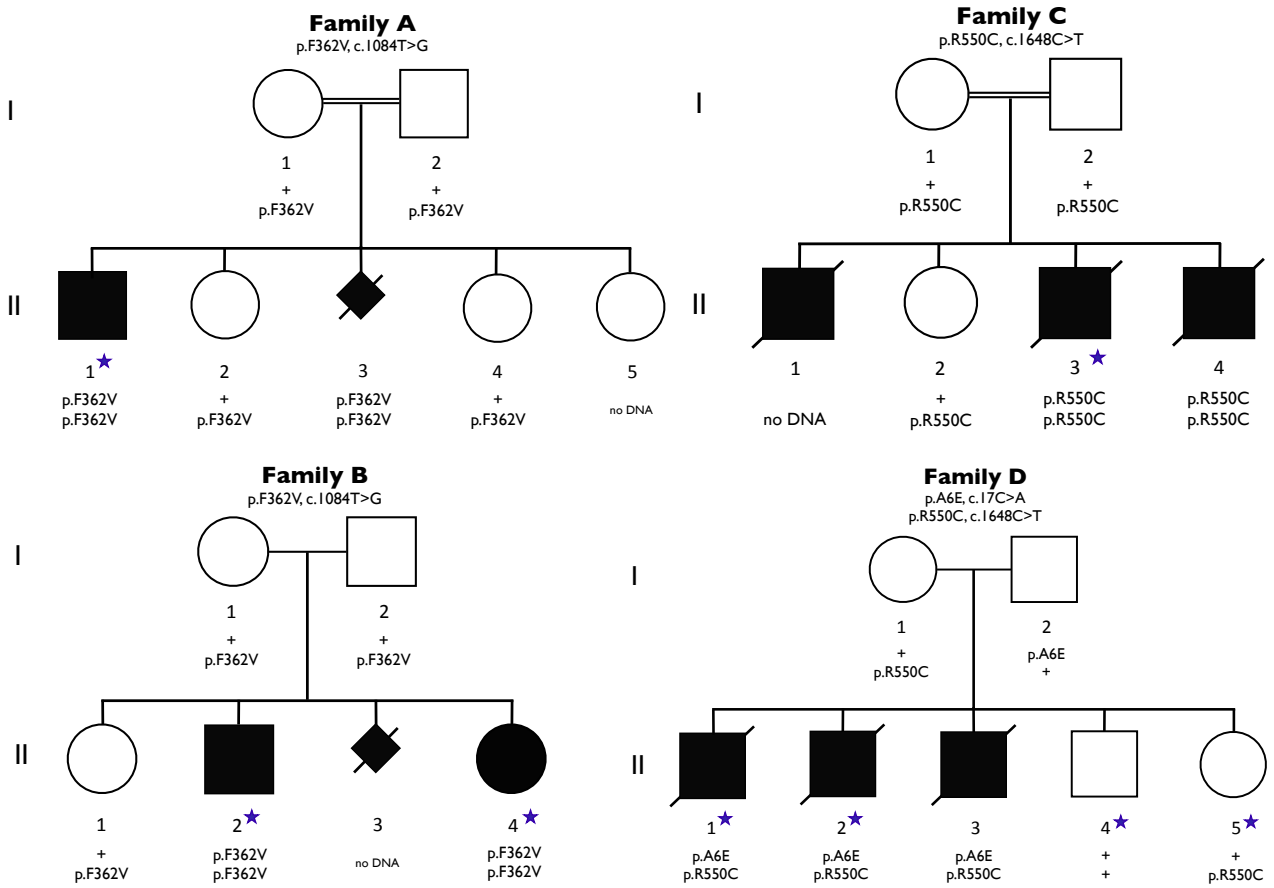


Figure 5. Four Families with ASNS Mutations.

Filled symbols represent known or presumed (in the case of aborted fetus) affected individuals. The individuals with a blue star were exome sequenced. Cosegregation of mutations is displayed for mutations in each of the four families.

Table 4. Clinical features of patients with mutations in *ASNS*

	Family A	Family B		Family C			Family D		
Consanguinity	Yes	No		Yes			No		
Ethnic origin	Iranian Jews	Iranian Jews		Bangladeshi			French Canadian		
Subjects	A.II.1	B.II.2	B.II.4	C.II.1	C.II.3	C.II.4	D.II.1	D.II.2	D.II.3
Genotype	p.F362V/p.F362V	p.F362V/p.F362V	p.F362V/p.F362V	Not determined	p.R550C/ p.R550C	p.R550C/ p.R550C	p.A6E/p.R550C	p.A6E/p.R550C	p.A6E/p.R550C
Gender	Male	Male	Female	Male	Male	Male	Male	Male	Male
Age	14 years	14 years	12 years	4 months†	3 months†	6 months†	9 days†	11 months†	12 months†
HC[1] at birth	31.5 cm (−2.5 SD)	31 cm (−3 SD)	31 cm (−2 SD/−3 SD)	30.5 cm (−3.5 SD)	33 cm (−1 SD/−2 SD)	32 cm (−2 SD)	31.5 cm (−2.5 SD)	31 cm (−3 SD)	28.5 cm (−1 SD/−2 SD)[2]
Developmental delay	Severe	Severe	Severe	Severe	Severe	Severe	Severe	Severe	Severe
Progressive microcephaly	Yes	Yes	Yes	Yes	Yes	Yes	NA	Yes	Yes
Epilepsy									
Age at onset	1 month	2 weeks	3 weeks	None	None	None	4 days	9 months[5]	8 days
Type of seizures	Spasms, tonic, myoclonic, GTC[3]	Spasms, tonic, myoclonic, GTC[3]	Spasms, tonic, myoclonic, GTC[3]	None	None	None	Tonic, orobuccal	Partial complex	Partial complex
EEG pattern	Hypsarrythmia, MISF[4]	Hypsarrythmia, MISF[4]	Hypsarrythmia, MISF[4]	Disorganized background	Disorganized background	Disorganized background	NA	Suppression bursts, MISF	Suppression bursts, MISF
Clinical examination									
Axial hypotonia	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Appendicular hypertonia	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hyperreflexia	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hyperekplexia	No	No	No	Yes	Yes	Yes	No	No	No
Brain MRI									
Decreased cerebral volume	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Decreased size of pons	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
simplified gyri	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes

†Deceased; [1]head circumference; [2]born at 33.5 weeks of gestation; [3]generalized tonic-clonic seizures; [4]multiple independent spike foci; [5]tremulous movements and abnormal EEG at 4 days; partial complex seizures with eye deviation and focal epileptic activity at 9 months.

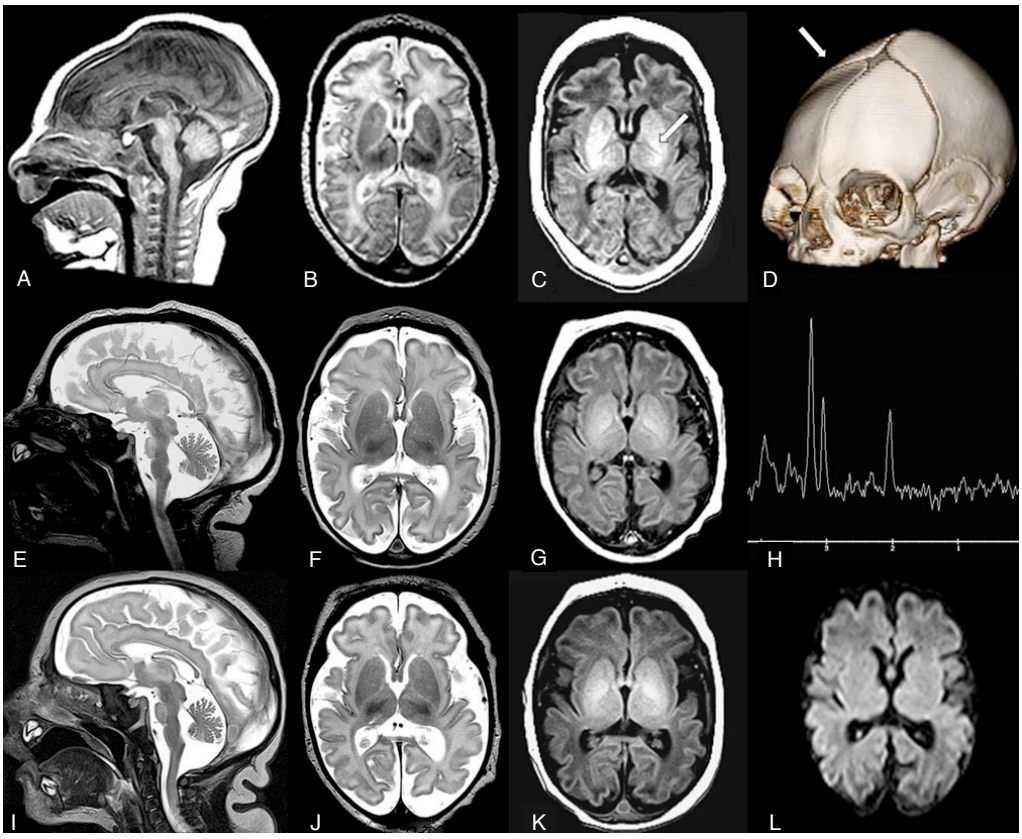


Figure 6. MRI Images from Family C.

(A–D) Sibling C.II.1. Sagittal T1W image (A) at 13 days of age reveals decreased size of pons. Axial T2W image (B) reveals prominent pericerebral fluid spaces surrounding the brain due to volume loss. White matter is diffusely higher in signal intensity than the cortical ribbon, suggesting delayed myelination. Axial T1W image (C) confirms lack of bright myelin stripe in the posterior limb of internal capsule (arrow). Three-dimensional computed tomography (D) at 2 months of age confirms ridge (arrow) from overlapping sutures, palpable at birth. DNA confirmation was not obtained in this infant. (E–H) Sibling C.II.3. Sagittal T2W image (E) performed at 6 days of age reveals decreased size of pons. Axial T2W image (F) prominent pericerebral fluid spaces surrounding the brain due to volume loss. Axial T1W image (G) lack of myelin stripe in posterior limb of the internal capsule (PLIC). Magnetic resonance spectroscopy (MRS) TE144 (H) performed in the basal ganglia is age appropriate (as it was in all three siblings). (I–L) Sibling C.II.4. Sagittal T2W image (I) at 1 day of age reveals decreased size of pons. Axial T2W image (J) pericerebral fluid prominence due to cerebral volume loss in a manner similar to siblings. Axial T1W image (K) shows lack of myelin stripe in PLIC. Diffusion-weighted imaging (DWI) (L) is normal (as it was in all three siblings).

Families A and B are unrelated but are both of Iranian Jewish ancestry. Targeted exonic regions were captured and sequenced in one affected individual from family A (A.II.1) and two from family B (B.II.2 and B.II.4). We focused on variants that were annotated as having a plausible impact on the function of the resulting gene product (e.g., missense, nonsense, splice site, intron-exon boundary, and coding-disrupting insertion-deletions (indels)). We compared patient exomes to control exomes sequenced in the same facility (n = 261, unrelated samples, not enriched for neurological disorders). Because families A and B belong to the same ethnic community and were the only similar cases identified in Israel to date, we postulated that the causal variant would be a founder mutation in this population shared among all affected individuals in these families. We therefore first focused on homozygous variants that were shared by both siblings in family B (Figure 5: B.II.2 and B.II.4) and that were uncommon in our control population (Table 5).

Table 5. Summary of rare variants shared in both patients from family B.
A summary of all rare variants, with a plausible impact on the function of the resulting gene product, shared in family B (found in both affected individuals II.2 and II.4). A rare variant was defined as having a predicted homozygote frequency of ≤ 3% in the population of 261 sequenced control genomes.

Functional category	Homozygous	Heterozygous
Essential splice site	1	2
Intron-exon boundary (SNV)	10	32
Intron-exon boundary (indel)	1	9
Non-synonymous coding	58	248
Stop gained	1	5
Coding-disrupted frameshift (indel)	1	8
Coding-disrupted other (indel)	0	2
Total	72	306

Since the incidence of this disorder is very low in the general population, we inspected only variants with a predicted frequency of ≤3% in our sequenced control genomes. We found 72 such variants, only three of which were absent in the control population (Table 6). Furthermore, only one of these three variants was also present in homozygous form in the patient from family A (Figure 5:A.II.1). This variant, located in the asparagine synthetase (*ASNS*) gene, causes a missense change (c.1084T > G) resulting in a phenylalanine to valine substitution at amino acid position 362 (p.F362V; NM_183356).

We also performed homozygosity mapping to identify shared regions of homozygosity between families A and B. We found that p.F362V lies in the largest homozygous region (~3 Mb) and found no additional candidates of interest in the homozygous regions.

Table 6. All rare homozygous functional variants shared in family B.

The 72 homozygous functional variants found in both II.2 and II.4 of family B. The variants are listed in ascending order based on the variant allele frequency in the sequenced control population. The bolded variant numbers denote the 9 variants that were also observed in the affected individual from family A (II.1).

Variant	Variant ID	Variant type	Gene symbol	Variant frequency	Variant	Variant ID	Variant type	Gene symbol	Variant frequency
1	7_97322654_C	SNV	ASNS	0.000	37	9_39075808_C	SNV	CNTNAP3	0.117
2	7_107186210_T	SNV	CBLL1	0.000	38	19_38355194_C	SNV	WDR88	0.119
3	7_107475774_C	SNV	LAMB4	0.000	39	19_59416242_A	SNV	LILRB3	0.119
4	7_106296855_T	SNV	PIK3CG	0.002	40	22_44146055_T	SNV	SMC1B	0.121
5	6_41721834_A	SNV	MDFI	0.004	41	X_1497953_T	SNV	ASMTL	0.121
6	12_51248317_A	SNV	KRT74	0.006	42	12_48631243_A	SNV	AQP2	0.123
7	7_106725874_C	SNV	COG5	0.008	43	16_2045401_T	SNV	TSC2	0.123
8	19_38307917_T	SNV	GPATCH1	0.010	**44**	15_19884068_C	SNV	OR4N4	0.123
9	7_104993051_A	SNV	RINT1	0.012	45	6_32740678_A	SNV	HLA-DQB1	0.125
10	7_98997458_G	SNV	ZNF655	0.015	46	22_29862960_T	SNV	PLA2G3	0.127
11	12_51152337_G	SNV	KRT6C	0.025	47	X_1500706_G	SNV	ASMTL	0.127
12	17_35550980_DEL_T	indel	CASC3	0.031	48	17_27649318_A	SNV	RHBDL3	0.129
13	16_2100974_G	SNV	PKD1	0.044	49	X_1506842_G	SNV	ASMTL	0.129
14	17_41416696_G	SNV	MAPT	0.046	50	11_77586662_A	SNV	USP35	0.131
15	7_107004256_A	SNV	DUS4L	0.048	51	22_30919090_T	SNV	RFPL2	0.131
16	6_24511434_T	SNV	MRS2L	0.050	**52**	19_59416243_A	SNV	LILRB3	0.135
17	18_12244956_A	SNV	CIDEA	0.054	53	16_2105631_C	SNV	PKD1	0.137
18	7_102361951_G	SNV	LRRC17	0.054	54	20_10551750_A	SNV	C20orf94	0.138
19	12_51838742_G	SNV	CSAD	0.056	55	3_41852418_C	SNV	ULK4	0.142
20	7_102452856_G	SNV	FBXL13	0.056	56	14_20569961_T	SNV	TPPP2	0.146
21	19_38209355_T	SNV	RHPN2	0.060	57	1_151037145_A	SNV	LCE1D	0.148
22	7_101885015_A	SNV	ALKBH4	0.062	**58**	12_51503256_A	SNV	KRT79	0.148
23	19_13901896_A	SNV	CC2D1A	0.079	59	1_37961428_A	SNV	EPHA10	0.160
24	16_2102362_G	SNV	PKD1	0.083	60	10_18868641_G	SNV	CACNB2	0.160
25	16_2095427_C	SNV	PKD1	0.087	**61**	7_29127129_T	SNV	CPVL	0.162
26	7_99655521_T	SNV	PVRIG	0.102	62	19_37859295_T	SNV	RGS9BP	0.162
27	1_37958776_T	SNV	EPHA10	0.104	63	15_41604696_A	SNV	MAP1A	0.163
28	7_100324690_C	SNV	UFSP1	0.104	64	16_2092388_G	SNV	PKD1	0.165
29	8_17202048_G	SNV	MTMR7	0.104	65	6_31432179_C	SNV	HLA-B	0.169
30	18_59805277_G	SNV	SERPINB8	0.104	**66**	15_19884261_G	SNV	OR4N4	0.169
31	3_62164229_A	SNV	PTPRG	0.108	67	1_75475262_A	SNV	SLC44A5	0.171
32	1_159963696_C	SNV	FCRLB	0.110	68	2_111315429_T	SNV	ACOXL	0.171
33	6_109430212_T	SNV	SESN1	0.110	69	11_77598578_G	SNV	USP35	0.173
34	X_152517671_INS_C	indel	FAM58A	0.112	70	3_47427122_A	SNV	PTPN23	0.175
35	12_51330534_T	SNV	KRT2	0.115	71	2_238675316_C	SNV	ESPNL	0.179
36	6_109587255_C	SNV	C6orf182	0.117	72	17_35133114_G	SNV	ERBB2	0.179

Family C is composed of three affected (C.II.1, C.II.3, and C.II.4) siblings and one healthy (C.II.2) sibling born to consanguineous parents of Bangladeshi origin (Figure 5). No DNA was available for the first affected child (C.II.1) who had the same clinical manifestations as his affected brothers. Homozygosity mapping showed that the two affected siblings share a total of eight homozygous regions that are >1 Mb in size (Table 7). Exome sequencing performed in one of the affected children (C.II.3) identified 856 rare protein or splice-altering variants (with a frequency ≤3% in 169 in-house unrelated exomes, 1,000 Genomes Project data set and data from the National Heart, Lung, and Blood Institute [NHLBI] Exome Sequencing Project [ESP]). These included three variants that map to the shared regions of homozygosity; the three variants were Sanger sequenced and all three variants were homozygous in both affected individuals. The parents and the unaffected sibling were heterozygous for two of these variants, whereas the other candidate variant was excluded from further consideration because it was found in a homozygous form in one of the parents. One of the remaining variants, c.1282G > A (p.D428N; NM_017460) in the CYP3A4 gene, is not predicted to affect protein function by SIFT or Polyphen-2 [92,93] and CYP3A4 encodes a component of cytochrome P450 (subfamily 3A, polypeptide 4), which is predominantly expressed in the liver. Thus, the CYP3A4 variant seemed unlikely to be responsible for this phenotype. The sole remaining variant in this family is c.1648C > T (p.R550C;

NM_183356) in ASNS, which is present in the largest region of homozygosity (35 Mb) shared by the two affected children (Table 8).

Table 7. Regions of shared homozygosity (>1Mb) between the affected individuals in Family C (II.3 and II.4).

Chr	Marker start	Marker end	Position start	Position end	Size (Mb)
1	rs3863722	rs272822	35,592,376	36,677,585	1.1
1	rs319950	rs9436447	49,091,687	50,590,732	1.5
3	rs3895736	rs3197999	48,658,467	49,721,532	1.1
3	rs9790150	rs2882429	133,723,028	139,835,865	6.1
4	rs10517277	rs6858830	33,466,803	34,500,856	1.0
5	rs10477652	rs10478752	123,179,118	126,087,378	2.9
7	rs11773446	rs7780168	95,944,485	131,291,677	35.3
12	rs1564121	rs12311684	86,138,548	89,435,177	3.3

Table 8. Exome sequencing variant filtering in family C.
*Rare variants were defined as those present at a frequency ≤3% in in-house control exomes (n=169), 1,000 genomes, and NHLBI Exome Sequencing Project.

Filter	Remaining variants
Coding /Splicing	20,439
Remove synonymous (non-splicing)	10,175
Rare*	856
Homozygous	25
In homozygous region	3
Segregating with the disease in the homozygous state	ASNS:NM_183356:exon14:c.C1648T:p.R550C CYP3A4:NM_017460:exon12:c.G1282A:p.D428N
Predicted damaging (SIFT or polyphen-2)	ASNS:NM_183356:exon14:c.C1648T:p.R550C

Family D is a nonconsanguineous French Canadian family, consisting of three affected (D.II.1, D.II.2, and D.II.3) and two unaffected (D.II.4 and D.II.5) siblings (Figure 5). Exome sequencing was performed in two affected (D.II.1 and D.II.2) and two unaffected siblings. In total, 237 rare protein or splice-altering variants were present in

both affected children (with a frequency ≤3% in 169 in-house unrelated exomes, 1,000 Genomes Project data set and data from the NHLBI ESP). We excluded from this list X-linked variants that were also present in the unaffected male sibling. We also excluded homozygous or possible compound heterozygous variants that were found in the same form in at least one unaffected sibling. Only two variants (c.1648C > T/p.R550C; c.17C > A/p.A6E; NM_183356), both in ASNS, remained after this filtering process (Table 9).

Table 9. Exome sequencing variant filtering in family D.
Rare variants were defined as those present at a frequency ≤3% in in-house control exomes (n=169), 1000 genomes, and NHLBI Exome Sequencing Project.

	Affected males		Unaffected male	Unaffected female
	II.1	II.2	II.4	II.5
Coding/Splicing	20,336	21,120	19,325	20,252
Remove synonymous (non-splicing)	10,021	10,491	9,793	10,363
Shared by II.1 and II.2	7,830			
Rare*	237			
Homozygous in II.1 & II.2 but *not* homozygous in II.4 or II.5	0			
Variants on chr X in II.1 and II.2 but not in II.4	0			
Genes with ≥2 rare variants in common to II.1 & II.2	14			
Genes with ≥2 rare variants in common to II.1 & II.2 but not in common with II.4 or II.5 (*compound heterozygous*).	ASNS (NM_183356) c.C1648T:p.R550C c.C17A:p.A6E			

Critically, in all four families there is complete cosegregation of the identified ASNS mutations/genotypes with disease (Figure 5). Sanger sequencing was used to validate all three mutations (Figure 7). For family D, Sanger sequencing also confirmed inheritance of each mutation from a different parent (compound heterozygote).

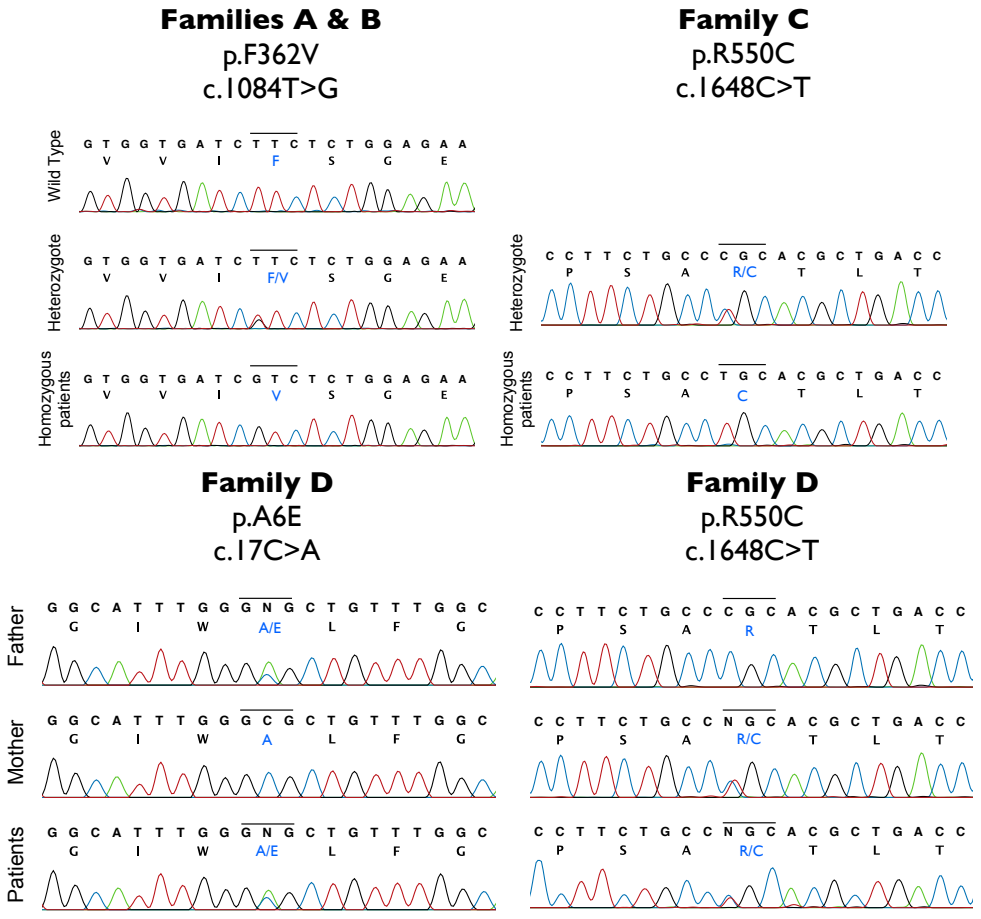


Figure 7. Sanger sequencing confirmation for all three ASNS mutations.

Table 10. Mutations identified in ASNS.

Family	Genotype	ASNS Modification	Nucleotide Change	Ethnicity	Frequency in Control In-house Exomes/Genomes	Frequency in Ancestry-Matched Controls	PolyPhen-2	SIFT
A	Homozygous	p.F362V	c.1084T>G	Iranian Jewish	0/261	1/80	Damaging (0.95)	Damaging (0.04)
B	Homozygous	p.F362V	c.1084T>G	Iranian Jewish	0/261	1/80	Damaging (0.95)	Damaging (0.04)
C	Homozygous	p.R550C	c.1648C>T	Bangladeshi	0/169	0/225	Damaging (1.00)	Damaging (0.01)
D	Compound heterozyous	p.R550C	c.1648C>T	French Canadian	0/169	0/300	Damaging (1.00)	Damaging (0.01)
		p.A6E	c.17C>A		0/169	0/300	Damaging (0.898)	Deleterious (0.02)

Nucleotide and amino acid positions are based on the NCBI Reference sequences NM_183356.3 and NP_899199.2, respectively.

All three missense mutations are predicted to damage the encoded asparagine synthetase protein by available computer algorithms (SIFT and PolyPhen-2) and all three mutations are absent in dbSNP135, the 1,000 Genomes Project data set, and data from the NHLBI ESP (Table 10). To better estimate the frequency of the p.F362V variant in unaffected individuals, we directly genotyped this locus in 1,160 additional controls and failed to detect the mutation. Finally, all three mutations were genotyped in ancestry-matched controls and all remained absent (Table 10), with the exception of p.F362V, which has an estimated carrier frequency of 0.0125 in Iranian Jews. Additionally, we used the sequence data to test for evidence of cryptic relatedness between the patient in family A and the affected siblings from family B and found no indication of elevated identity by descent beyond what is expected for unrelated genomes (data not shown). We also tested whether the p.F362V ASNS variant is found on a common haplotype in all affected individuals of Iranian Jewish origin. Indeed, the ASNS variant was found on the same 1.2 Mb haplotype in both families and this haplotype was very rare (0.8%) in 261 sequenced controls (Table 11). This observation is consistent with a single founder origin for p.F362V and subsequent transmission along with the same extended haplotype.

Table 11. The shared F362V ASNS haplotype in families A and B.
This haplotype is defined by 16 variants and the first SNV is the p.F362V variant.

SNV	Coordinate	Alleles
1	chr7:97322654	C/C
2	chr7:97325666	A/A
3	chr7:97326505	T/T
4	chr7:97621997	C/C
5	chr7:97654263	T/T
6	chr7:97660051	A/A
7	chr7:97660146	A/A
8	chr7:97760787	T/T
9	chr7:97760811	A/A
10	chr7:97771537	T/T
11	chr7:97782680	A/A
12	chr7:98298760	A/A
13	chr7:98338892	T/T
14	chr7:98396816	C/C
15	chr7:98487987	C/C
16	chr7:98488035	C/C

Interestingly, the mutation p.R550C was found in two families of different ethnic backgrounds. This mutation was associated with different haplotypes in each of these families, suggesting that it arose independently. It should be noted that p.R550C corresponds to a CpG site, which is associated with a higher mutation rate [94], possibly explaining the recurrence of this rare mutation in different populations.

2.3.2 Functional impact of the nonsynonymous mutations

To test the effect of the identified mutations on ASNS mRNA and protein levels, we generated full-length mutant cDNA constructs (p.A6E, p.F362V, and p.R550C) using PCR-mediated site-directed mutagenesis (Figure 3). We then transfected both wild-type and mutant alleles into HEK293 and COS-7 cells and found similarly robust levels of

expression of the mRNA corresponding to wild-type and all three mutant alleles (Figure 8). This result indicates that these mutations do not overtly affect mRNA levels, suggesting that they do not influence mRNA stability. For the p.F362V mutation, we also compared wild-type and mutant full-length transcripts, from the patient fibroblasts, to detect any differences in alternative splicing or exon skipping and found no evidence for alternately spliced transcripts (data not shown).

We used two approaches to detect the ASNS protein in transfected cells. First, we used an antibody to human ASNS, but this antibody cross-reacted to produce nonspecific bands (Figure 9). Thus we also used C-terminal FLAG-tagged forms of ASNS to detect the wild-type and all three mutant forms of ASNS (Figure 4) using the anti-FLAG antibody. We found that while high levels of the wild-type protein were easily detected, a dramatic reduction in protein abundance was seen in the HEK293 cells expressing the p.A6E or p.F362V mutant allele. In contrast, cells expressing the p.R550C mutant allele had an increased level of protein abundance compared to wild-type (Figure 8(B)). Consistent with the former observation, a dramatic reduction in ASNS abundance was observed in patient fibroblasts from individual II.1 in family A, harboring the p.F362V allele (Figure 9).

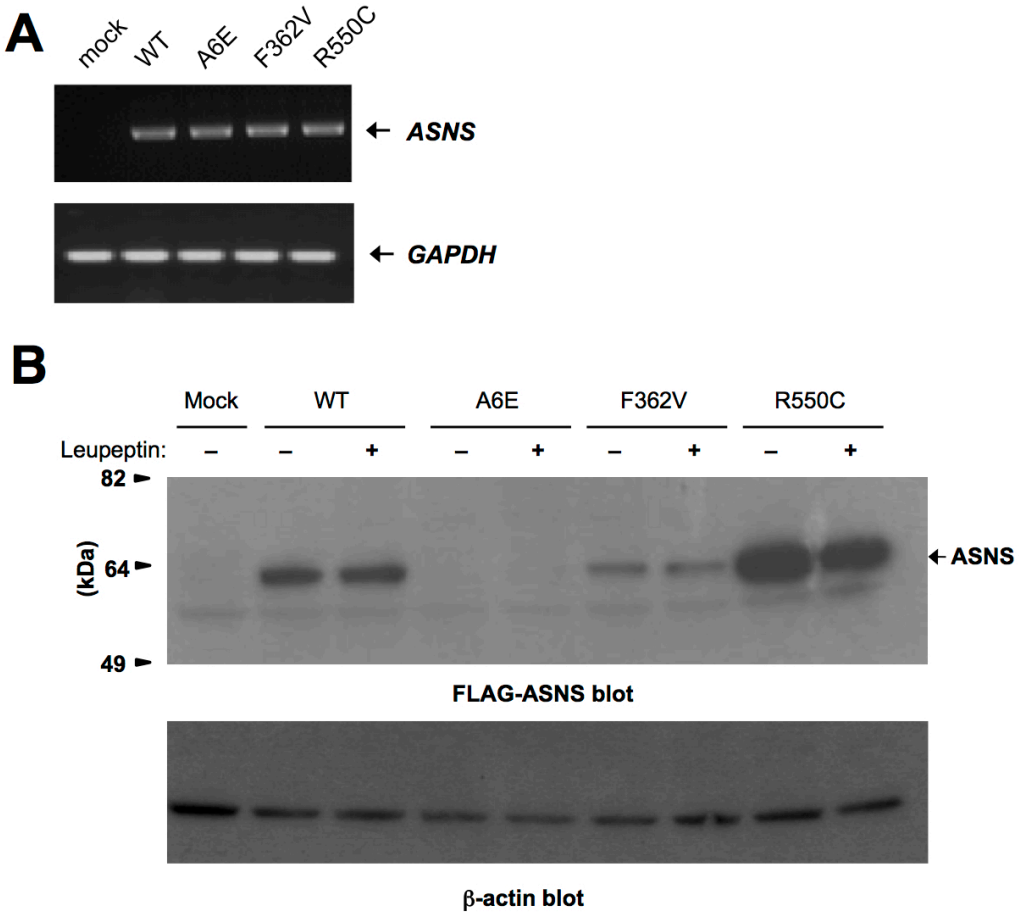


Figure 8. Functional impact of ASNS mutations.

(A) RT-PCR to detect ASNS mRNA expression in COS-7 cells transfected with empty, wild-type, or mutant vectors. (B) Western blots detecting ASNS-FLAG protein abundance, with or without Leupeptin treatment, in the HEK293 transfectants using an anti-FLAG antibody. β-actin was used as a loading control.

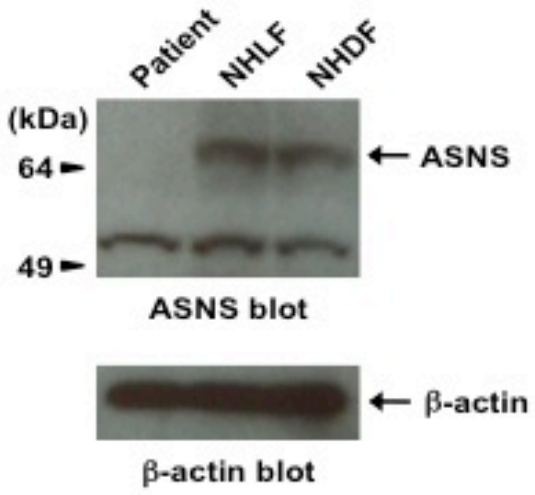


Figure 9. ASNS levels in patient fibroblast cells.
ASNS in a patient with the F362V mutation (Family A: II.1) was detected using an anti-ASNS antibody (also detects a nonspecific band). NHDF and NHLF are human fibroblast cell lines. The protein in the patient cells is depleted as was seen with the *in vitro* mutagenesis studies.

This pattern of protein abundance was also observed in COS-7 cells transfected with empty, wild-type, or mutant vectors (Figure 10). These results suggest that these mutations impair ASNS gene function by either reducing protein expression (p.A6E or p.F362V) or reducing functional performance (p.R550C). The mechanism through which the R550C mutation reduces activity remains to be elucidated, but the clinical similarity in presentation of patients suggests that all mutations are loss of function mutations.

We also asked whether these mutations destabilize the protein, targeting it for degradation. We blocked both the ubiquitin-proteasome and the macroautophagy pathways, but neither of these altered ASNS protein abundance (data not shown). We also used Leupeptin to inhibit lysosomal-dependent degradation and this also failed to

rescue the p.A6E or p.F362V mutant proteins to wild-type levels (Figure 8(B), Figure 10), although some experiments did show a trend toward rescue (data not shown).

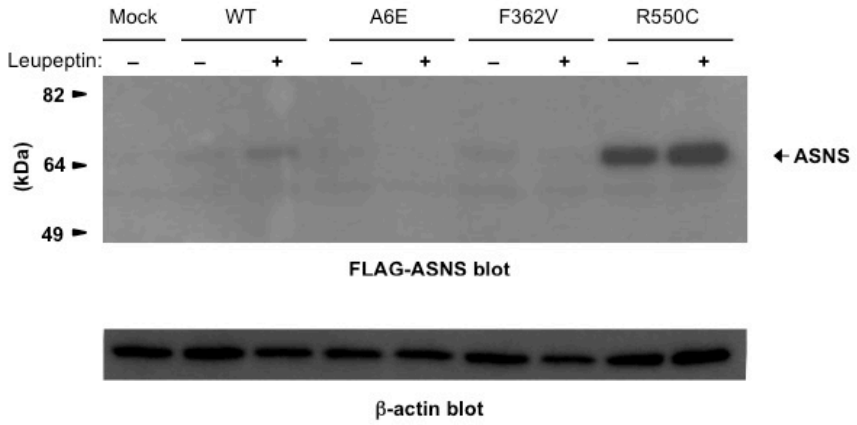


Figure 10. ASNS levels in COS-7 cells transfected with empty, wild-type, or mutant vectors.

Western Blots detecting ASNS-FLAG protein abundance, with or without Leupeptin treatment, in COS-7 transfectants using an anti-FLAG antibody. β-actin was used as a loading control.

ASNS encodes the glutamine-dependent asparagine synthetase enzyme (EC 6.3.5.4), which catalyzes ammonia transfer from glutamine to aspartic acid via a β-aspartyl-AMP intermediate. Concordant with this biochemical function, we found that the levels of asparagine were decreased in at least two affected individuals (C.II.3 and D.II.1), whereas glutamine and aspartic acid, both precursors in the ASNS-catalyzed synthesis of asparagine, were mildly elevated in the patients from family B (Table 12).

These findings are consistent with our in vitro functional studies, emphasizing that the identified mutations have phenotypic consequences.

Table 12. Measurements of amino acid concentrations in patient blood and urine.

	Plasma Amino Acid Levels			Urine Amino Acid Levels		
	Glutamine	Aspartate	Asparagine	Glutamine	Aspartate	Asparagine
B.II.2	1,250 (254–823 μmol/l)	18 (1–24 μmol/l)	57 (23–112 μmol/l)	382 (20–112 μmol/mmol creatinine)	36 (1–10 μmol/mmol creatinine)	21 (0–24 μmol/mmol creatinine)
B.II.4	1,149 (254–823 μmol/l)	2 (1–24 μmol/l)	49 (23–112 μmol/l)	57 (20–112 μmol/mmol creatinine)	40 (1–10 μmol/mmol creatinine)	4 (0–24 μmol/mmol creatinine)
C.II.1	NA	7 (17–21 μmol/l)	NA	NA	NA	NA
C.II.3	NA	NA	12 (16–21 μmol/l)	NA	NA	NA
C.II.4	NA	12 (0–20 μmol/l)	NA	NA	NA	NA
D.II.1	439 (474–736 μmol/l)	7 (4–18 μmol/l)	11 (31–56 μmol/l)	NA	NA	NA
D.II.2	668 (474–736 μmol/l)	9 (4–18 μmol/l)	55 (31–56 μmol/l)	NA	NA	NA

Data were not available for all fields in all patients (NA). The reference concentrations are indicated within parentheses.

The mutated amino acid residues in ASNS are located within regions of high sequence conservation among orthologs, from bacterium to man (Figure 11(A)), indicating that these amino acids are likely to be critical for protein function. This is further supported by the inferred positions of the human ASNS mutations in the folded bacterial ortholog (Figure 11(B)). The inferred p.A6E position is located in the N-terminal domain (which is responsible for glutamine hydrolysis), which faces a pocket identified as the glutamine binding site [95]. The inferred p.F362V position is located on the C-terminal domain, near the AMP binding site. The inferred p.R550C position is located at the distal end of the C-terminus which is disordered in the E. coli ortholog and thus the crystallographic structure is not available. However, this region is in the interface between the N-terminal and C-terminal domains and it might play a role in binding the aspartate substrate [95].

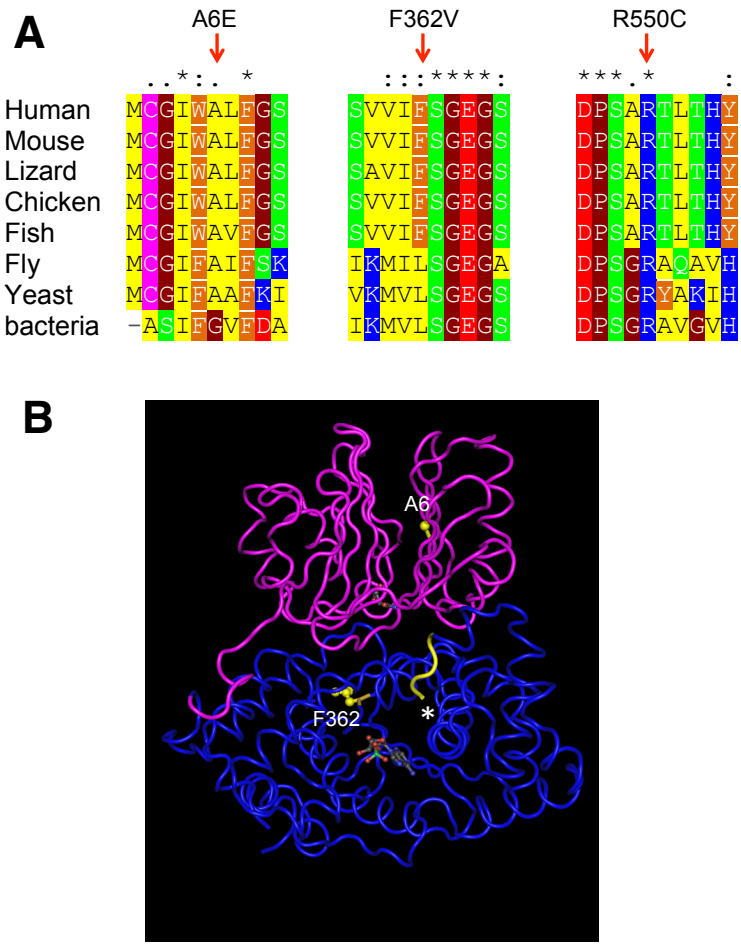


Figure 11. Location and conservation of mutated residues in ASNS.
(A) A multiple alignment of human ASNS and selected orthologs. Only the regions harboring mutations are shown. On top: ''*'' represents identical position; '':'' represents conserved substitutions; ''.'' represents semiconserved substitution. Color code is for amino acid type: red, negative; blue, positive; green, polar; yellow, aliphatic; orange, aromatic; brown, helix breaker. **(B)** The structure of *E. coli* glutamine-dependent asparagine synthetase B (protein databank ID 1CT9), a bacterial ortholog of human ASNS (A). The N-terminal glutamine aminotransferase domain is colored in pink, and the C-terminal asparagine synthase domain is colored in blue. The residues inferred to be equivalent to the mutated A6E and F362V are highlighted in yellow and the approximate location of the mutated R550C is shown with an asterisk (crystallography could not be determined for the distal end of C-terminal domain). Also, the AMP and glutamine molecules are shown in space-filling style.

2.3.3 ASNS expression in the brain

Cells are capable of both nutritional intake and endogenous synthesis of asparagine, suggesting that ASNS may be dispensable, and raising the question of how loss of ASNS protein or its dysfunction results in a severe, tissue-specific phenotype. *ASNS* is under complex transcriptional regulation [96-99] and is expressed at low levels in most tissues but is highly expressed in the adult brain, with evidence for a brain-specific splice variant(s) (Figure 12) [100].

Consistent with this expression pattern, *Asns* is expressed in the adult mouse brain (Allen Brain Atlas). In situ hybridization shows that *Asns* is also expressed in the developing embryonic mouse brain (E14.5), is particularly enriched in the cortical plate where the neurons reside, and is also expressed in the ventricular and subventricular zone layers (VZ and SVZ) lining the ventricles of the cerebral cortex, where neural progenitors reside (Figure 13) [101]. RNA-seq of E14.5 cortices also confirms this pattern of expression in the cortical plate and VZ [102]. This expression pattern is similar to that of known microcephaly genes [103,104], consistent with a role for *Asns* in cortical development and brain size.

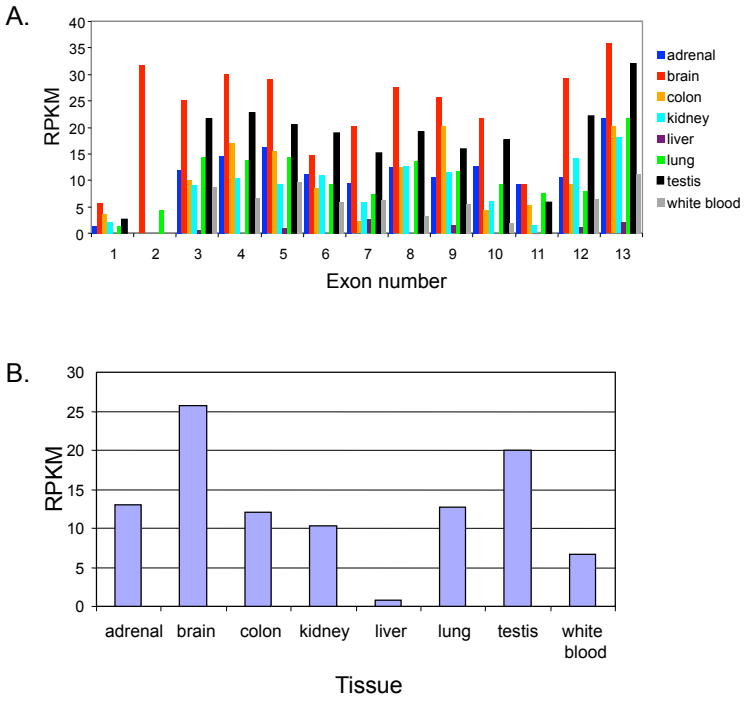


Figure 12. ASNS expression in different tissues.

Expression profile of *ASNS* gene (ENST00000394308) using RNA-seq data, extracted from the Human Body Map 2.0 Project of Illumina (downloadable from Integrative Genomics Viewer (IGV) at http://www.broadinstitute.org/igv/). In this database, the RNA from each tissue was obtained from a single individual. Extracted reads were aligned to the human genome (hg19) using Bowtie [105], and BEDTools [106] was applied for counting the number of reads per exon. Data were normalized to reads per kilobase of exon model per million mapped reads (RPKM) values. **(A)** Normalized expression per exon, showing an indication of brain-specific expression of exon 2. **(B)** Normalized expression for the whole gene. The finding that *ASNS* is more highly expressed in brain and testis is in agreement with Horowitz *et al.* [107].

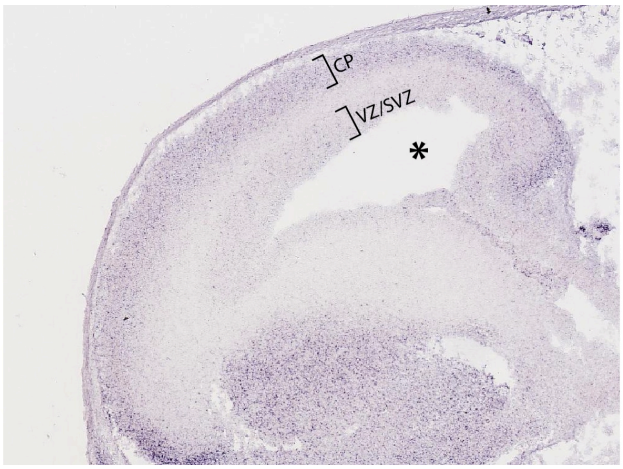


Figure 13. Asns expression in the developing mouse brain.
Asns mRNA *in situ* hybridization of the developing mouse brain (sagittal section) at E14.5. *Asns expression* is particularly enriched in the cortical plate (CP) and ventricular and sub-ventricular zone layers (VZ/SVZ). The asterisk is labeling the ventricle of the developing mouse cerebral cortex. *In situ* image provided by http://www.eurexpress.org/ (euxassay_004453).

2.3.4 Asns gene-trap mice

We obtained genetically modified mice from EUCOMM in which the *Asns* genomic locus contains a gene-trap insertion in intron 2 (ENSMUST00000115542) containing a splice acceptor site and the LacZ gene (B6NTac;B6N-Asns$^{tm1a(EUCOMM)Wtsi}$/H, termed *Asns*$^{-/-}$; Figure 14).

Gene traps are frequently hypomorphs, as there can be splicing that skips over the gene-trap cassette, but the degree to which this splicing occurs is construct and gene specific [108]. To determine the extent of *Asns* expression in the *Asns* gene-trap mice, we performed a comprehensive *Asns* mRNA analysis in the brains (cerebral hemispheres) of adult *Asns*$^{+/+}$, *Asns*$^{+/-}$, and *Asns*$^{-/-}$ mice (3 to 4 months of age). First, RT-PCR was used to

semiquantitatively assess the existence of *Asns* mRNA transcripts. RT-PCR spanning from the second exon to the last exon (exon 12) revealed a single band the size of the expected wild-type *Asns* transcript in all three genotypes (Figure 15). These bands were gel purified and Sanger sequenced, which confirmed that the wild-type transcript was present in all three genotypes and no aberrant splicing events were observed. Additionally, *Gapdh* RT-PCR was performed as an internal control and the homozygous mice show a decreased abundance of the wild-type *Asns* transcript compared to the wild-type mice (Figure 15).

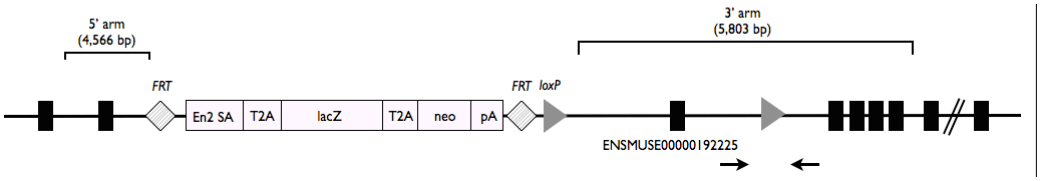


Figure 14. Asns gene-trap construct.
Genetically modified mice from EMMA (http://www.emmanet.org/strains.php). This diagram shows the complete construct including the with a splice acceptor site, after exon 2, for the lacZ gene (B6NTac;B6N-Asns$^{tm1a(EUCOMM)Wtsi}$/H); custom genotyping primer locations shown as black arrows.

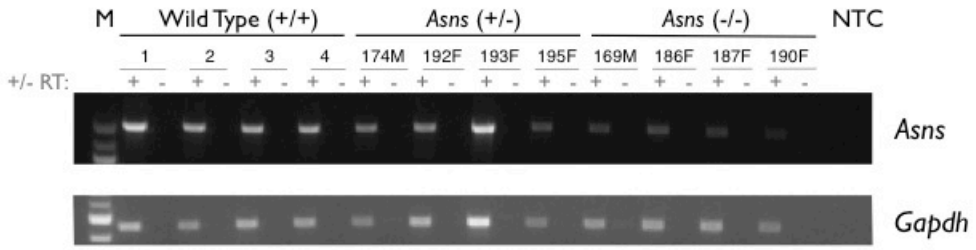


Figure 15. Asns adult mouse brain semi-quantitative RT-PCR.
RT-PCR with low cDNA input (250ng) was used to assess the existence of *Asns* mRNA transcripts (spanning from exon 2 to exon 12) and *Gapdh* mRNA transcripts semi-quantitatively.

To quantify this difference, we performed quantitative real-time TaqMan PCR using a probe spanning exons seven and eight. The mRNA levels were significantly different between the three genotypes (one-way ANOVA; $p < 0.00001$) (Figure 16). A post hoc two-tailed t test revealed that both mutant genotypes were significantly different from wild-type mRNA levels ($P_{Asns(+/+)\text{-}Asns(+/-)} = 0.00001$, $^{*}P_{Asns(+/+)\text{-}Asns(-/-)} < 0.00001$) and significantly different from each other ($^{*}P_{Asns(+/-)\text{-}Asns(-/-)} = 0.00083$). We were unable to determine whether there was differential expression of the Asns protein, due to lack of quality and specific mouse anti-Asns antibodies. These data demonstrate that the *Asns* gene-trap mouse is a hypomorph with ~20% of the normal level of *Asns* mRNA being expressed. Given that two of the human mutations lead to decreased protein expression, this mouse provides a reliable model for this phenotype.

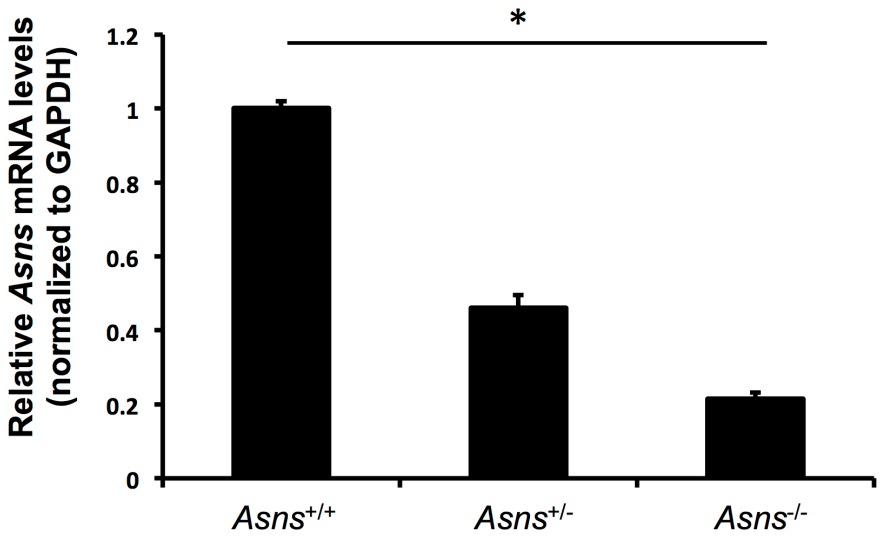


Figure 16. Detection of Asns mRNA (qRT-PCR) in the mouse brain.
Four mice of each genotype were used (all 3–4 months of age). The expression differences between the three genotypes are significant (one-way ANOVA test *p < 0.00001). A post hoc two-tailed t test revealed both mutant genotypes were significantly different from wild- type expression ($P_{Asns(+/+)\text{-}Asns(+/-)} = 0.00001$, $*P_{Asns(+/+)\text{-}Asns(-/-)} < 0.00001$) and significantly different from each other ($*P_{Asns(+/-)\text{-}Asns(-/-)} = 0.00083$). Error bars represent SEM.

We next analyzed the brains of *Asns*$^{-/-}$ and control (*Asns*$^{+/+}$ or *Asns*$^{+/-}$) littermates from embryos and adults. We obtained coronal sections from postnatal day (P) 0 brains and measured cortical area, cortical thickness, and lateral ventricle area for each mouse using rostral-caudal-matched sections (using anatomical landmarks). We found that the cortical thickness and area of the *Asns*$^{-/-}$ brains were, on average, ~14% thinner and ~5% smaller than their control littermates, respectively. Additionally, the lateral ventricle area in the *Asns*$^{-/-}$ mice was significantly larger than their control littermates ($p = 0.019$; Figure 17).

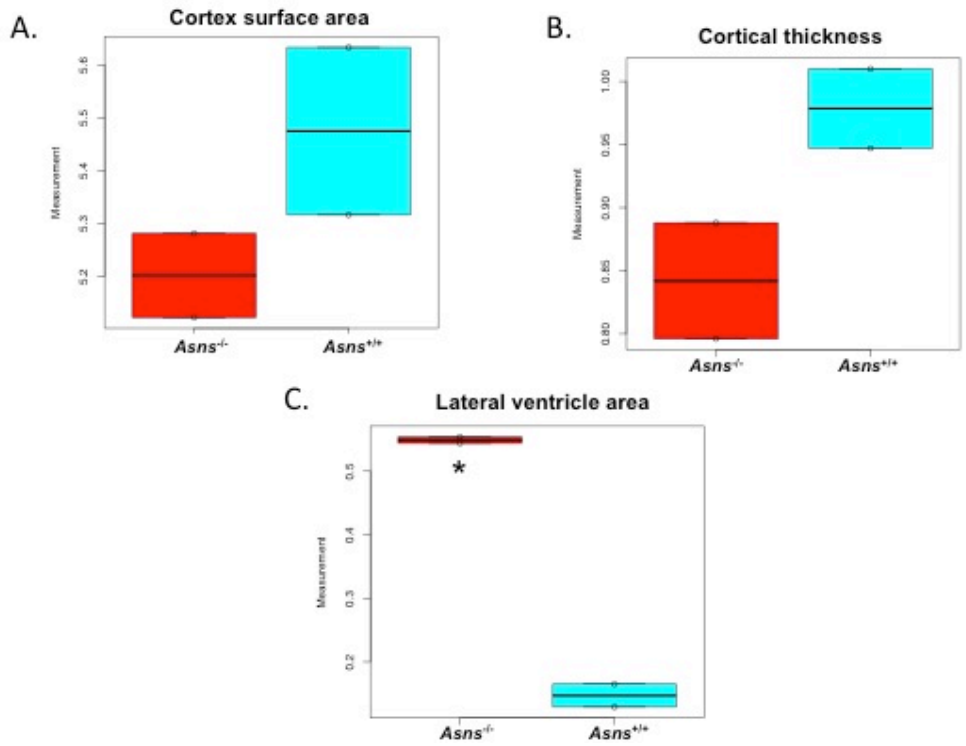


Figure 17. Postnatal day zero (P0) mouse brain measurements. Measurements were taken using two homozygous mutant mouse and two age matched (P0) wild type littermates. **(A)** Surface area of the mouse cortex (right hemisphere only) **(B)** cortical thickness and **(C)** Lateral ventricle area of right hemisphere; * indicates a significant difference by an unpaired t-test (p=0.019). Measurement units are arbitrary.

Due to the progressive nature of the human disorder, we next evaluated whether adults showed exacerbated brain defects. We generated paraffin-embedded coronal sections from P84 brains of *Asns*$^{-/-}$ and *Asns*$^{+/-}$ littermates (three of each genotype) (representative sections shown in Figure 18). The use of heterozygous animals was considered suitable because human carriers of *ASNS* mutations remain unaffected. We analyzed rostral-caudal-matched sections (using anatomical landmarks) from each animal for several parameters. Measurement of the cortical surface area, using methods

described by Pulvers and colleagues [109], showed an ~8% reduction in cortical surface area of *Asns*$^{-/-}$ mice (Figure 19(A)). A similar reduction (~5%) was observed in the whole-brain surface area of *Asns*$^{-/-}$ mice (Figure 19B). We also observed that the *Asns*$^{-/-}$ brains had increased lateral ventricles (~95%) relative to control brains (Figure 19(C)). Importantly, the cortical thickness of the *Asns*$^{-/-}$ mice was significantly reduced compared to the *Asns*$^{+/-}$ mice ($p = 0.022$; Figure 19(D)).

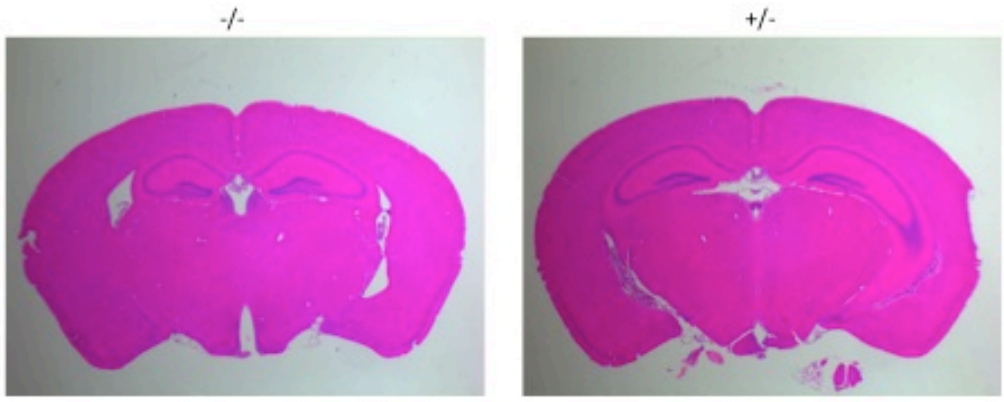


Figure 18. Asns adult mouse brain sections.
Coronal sections from one homozygous mutant mouse (-/-) and an age matched (P84) heterozygous littermate (+/-).

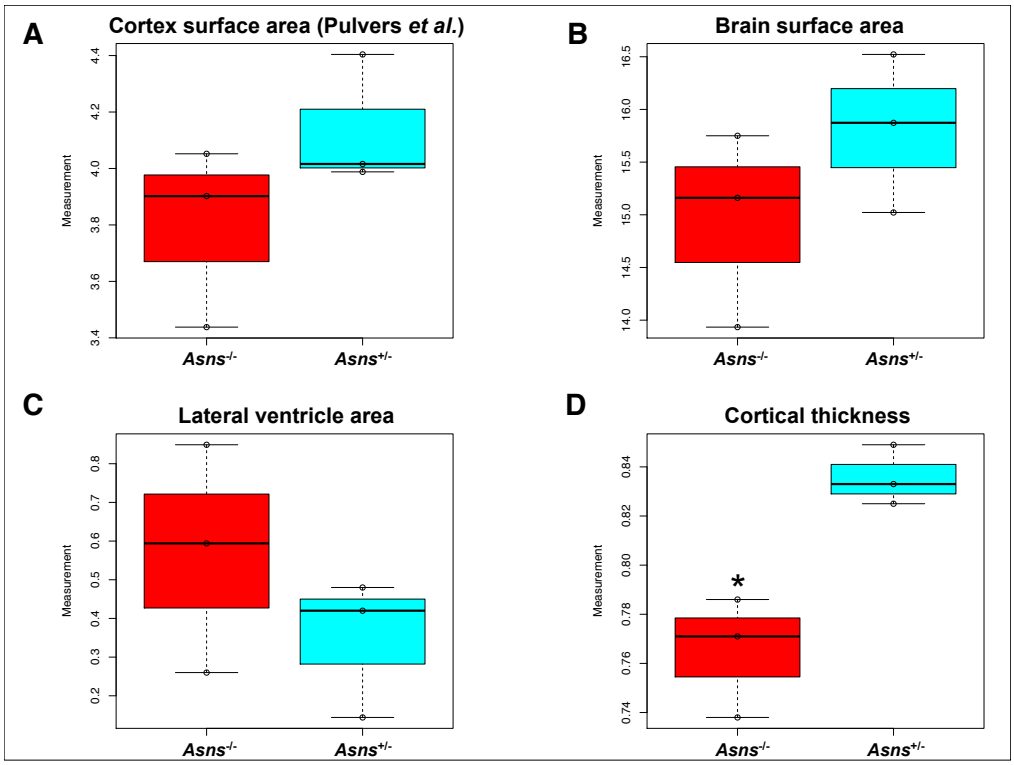


Figure 19. Structural brain abnormalities in Asns-deficient mice.
Measurements of adult mouse brain (P84) coronal sections, analyzed using ImageJ software [110], comparing three homozygous mutants to heterozygous littermates. **(A)** Cortical surface area as measured in Pulvers *et al.* [109]. **(B)** Surface area of the left hemisphere of the brain section, **(C)** surface area of the lateral ventricle in the left hemisphere of the brain section, and **(D)** cortical thickness measured from the edge of the hippocampus to the outer cortex. Asterisk (*) indicates a significant difference by an unpaired t test ($p = 0.022$). Error bars represent the range of the observed data.

Asns$^{+/-}$ and *Asns*$^{-/-}$ mice were assessed with age-matched B6NTac control animals in four behavioral assays. We found no genotype-associated differences in spontaneous locomotor activity, performance on the rotorod, or anxiety-like behavior in the light-dark emergence test; however, *Asns*$^{+/-}$ mice were deficient and *Asns*$^{-/-}$ mice were severely impaired in short- and long-term memory (Figure 20). The novel object recognition task evaluates short- (STM) and long-term memory (LTM) [111]. In the STM test, Asns+/- and

Asns-/- mice showed no preference for either the novel or familiar object (Figure 20(H)). In the test for LTM, the $Asns^{-/-}$ mice failed to demonstrate a preference for either object. By comparison, CONT and $Asns^{+/-}$ animals preferred the novel object; however, LTM was still reduced in the $Asns^{-/-}$ mice. It should be emphasized that the deficits in STM or LTM by the mutants were not related to the total time spent with the objects or to the total numbers of object contacts (Figure 20 (F-G)). Collectively, these data indicate that $Asns^{-/-}$ mice are impaired in both STM and LTM. By contrast, there is some sparing of LTM in the $Asns^{+/-}$ animals, suggesting that memory consolidation is delayed in these mutants.

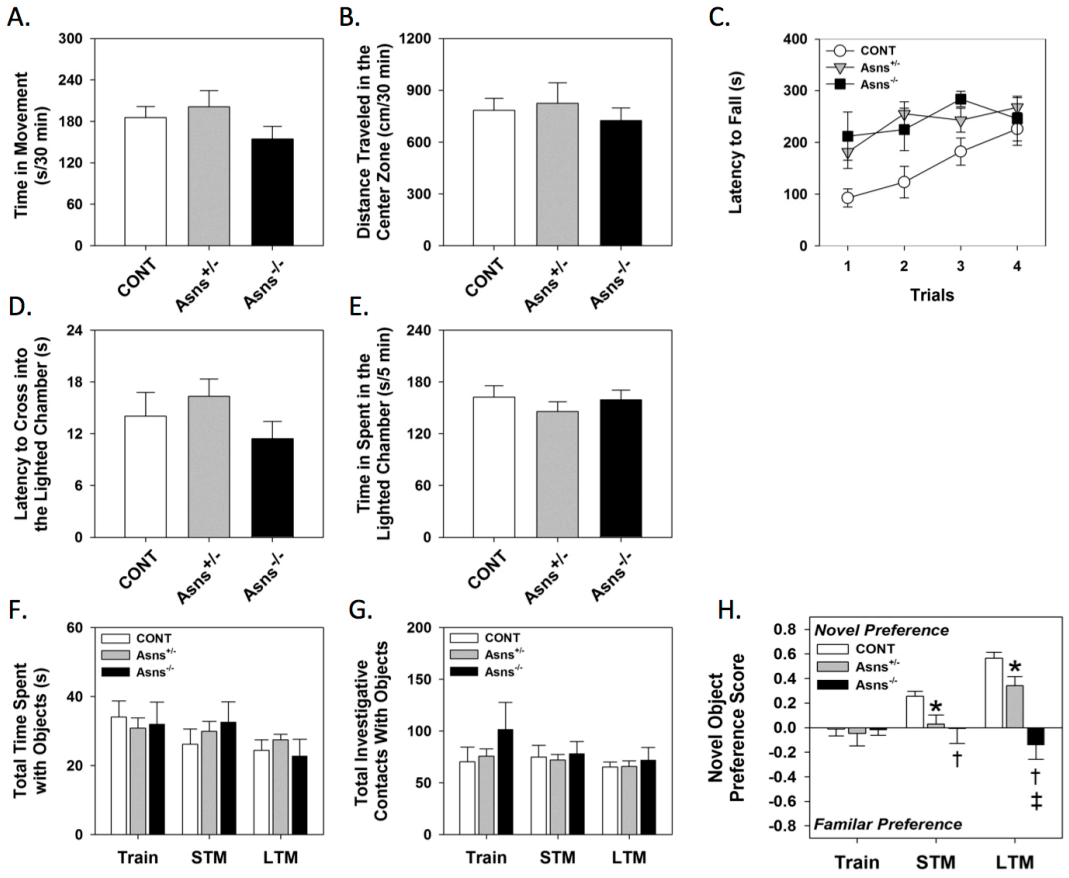


Figure 20. Behavioral analyses of *Asns* mice.

(A-B) Spontaneous locomotor activity in the open field over 30 min for B6NTac control (CONT), Asns$^{+/-}$, and Asns$^{-/-}$ mice: time in movement in the open field (A) and distance traveled in the center zone of the open field (B). **(C)** Latency to fall from the

rotorod across four trials. **(D-E)** Light-dark emergence test: latency to enter the lighted chamber (D) and time spent in the lighted chamber (E). **(F-G)** Novel object recognition task as assessments during training (Train), short-term (STM), and long-term memory tests: total time spent with objects (F), total investigative contacts with objects (G), and preference scores from the comparison between novel and familiar objects (H). *p<0.05 (CONT versus $Asns^{+/-}$ mice), †p<0.05, (CONT versus $Asns^{-/-}$mice), ǂp<0.05 ($Asns^{+/-}$ versus $Asns^{-/-}$ mice).

Careful observations of mice during behavioral testing revealed no evidence of seizure activity. To examine the possibility that *Asns*$^{-/-}$ mice might display epileptiform activity, we conducted prolonged video electroencephalogram (EEG) recordings in chronically implanted *Asns*$^{-/-}$ mice (n = 2) and a wild-type (WT) control (n = 1). Neither behavioral nor electrographic seizures were detected in *Asns*$^{-/-}$ mice or the WT controls.

Taken together, these data indicate that this Asns mouse model recapitulates the human brain phenotype, particularly in the reduced cortical thickness and increased lateral ventricle area.

2.4 Discussion

We report that mutations in asparagine synthetase (*ASNS*) cause a distinct neurodevelopmental disorder characterized by congenital microcephaly, profound intellectual disability, and progressive cerebral atrophy. We found that two of these mutations reduce the abundance of the protein. Finally, we have shown that disrupting this gene in mice creates a model that mimics aspects of the human phenotype, including structural brain abnormalities and learning deficits, albeit with what appears to be a generally milder presentation than observed in humans.

Studies performed on cancer cells showed that asparagine depletion affects cell proliferation and survival (reviewed in [99]). This is classically illustrated by the effect of asparaginase administration in childhood acute lymphoblastic leukemia. Asparaginase delivery to the blood- stream results in asparagine depletion causing a rapid efflux of

cellular asparagine, which is also destroyed. Most cells express sufficient ASNS to counteract this asparagine starvation and survive, but not leukemic cells. Similarly, loss of ASNS activity in thermosensitive mutant BHK cells leads to cell-cycle arrest as a consequence of a depletion of cellular asparagine [97,98].

During development, *Asns* is expressed in regions where both neural progenitors and postmitotic neurons are present, suggesting that it may function in either or both of these populations. A subset of the brains from our subjects had simplified gyri. Similar features were found in the mutant mice, which showed decreased cortical thickness and enlarged lateral ventricles. These structural abnormalities could be caused in part by aberrations in neural progenitor proliferation during development, resulting from decreased asparagine levels. Asparagine depletion could also cause increased cell death in postmitotic neurons or glial cells, contributing to the progressive atrophy of the brain observed in our subjects.

Strikingly, ASNS deficiency causes severe neurological impairment, without any involvement of peripheral tissues. The concentration of asparagine in the cerebrospinal fluid (CSF) of humans is only ~10% of the concentration found in plasma [112]. The poor transport of asparagine across the blood-brain barrier suggests that the brain depends on local *de novo* synthesis, explaining why the phenotype is essentially neurological.

In addition to ID, a subset of our patients presented with features of hyperexcitability (including epilepsy and hyperekplexia). These features suggest a

mechanism that is consistent with the accumulation of aspartate/glutamate in the brain, resulting in enhanced excitability and neuronal damage. While seizures in the patients could reflect enhanced excitability, these could also be secondary to the structural effects of altered proliferation. We cannot exclude the possibility that multiple mechanisms may be contributing to the observed human phenotype. Further analyses of animal and cellular models will help to elucidate the function of ASNS in normal brain development.

Of particular interest is the observation that *Asns* hypomorphic mice appear to have a milder phenotype than the humans with regards to more modest structural effects on the brain and no evidence of seizures. The ratio for the concentration of asparagine in the CSF to plasma in rats (0.26) [113] appears to be slightly elevated compared to that of humans (0.081 [114] to 0.118 [112]). Assuming that the CSF/plasma ratio is similar in mouse and rat, this suggests that the concentration of asparagine is increased in the CSF and interstitial fluid (ISF) of mouse/rat as compared to humans. Thus, asparagine may be more readily available to the *Asns*$^{-/-}$ mice due to some physiological difference between humans and mice, such as transport at the blood-brain barrier. Alternately, it is possible that low levels of *Asns* expression in these mice result in a less severe phenotype. It will be of great interest to compare the hypomorph to a complete *Asns* null animal, which may show an even more dramatic phenotype.

With this report, ASNS deficiency becomes the third example of a recently recognized group of conditions resulting from the inability to synthesize a nonessential amino acid. These conditions all feature severe congenital encephalopathy with microcephaly. The others are glutamine synthetase deficiency [115] and the serine biosynthetic disorders [116]. Although knowledge of ASNS deficiency and of other inborn errors of nonessential amino acid synthesis is incomplete, general considerations regarding diagnosis, disease mechanism, and treatment are in order. In almost every respect, the clinical approach to these diseases is predicted to be the opposite of that recommended for classical aminoacidopathies, which are caused by deficient breakdown of essential amino acids.

Strikingly, every diagnosis of ASNS deficiency was made by molecular genetics, despite extensive previous evaluation of patients that in several cases included amino acid chromatography of plasma and CSF. Why was ASNS deficiency not suspected on these grounds? The answer may lie in a combination of technical considerations and biology. Compared to most amino acids, the normal levels of asparagine are low, both in plasma (e.g., 50.7 ± 17.7 mmol/l, in children 0–3 years old) and CSF (e.g., 4.0 ± 2.9 mmol/l) [112,114]. For many reasons, low levels of a metabolite may be less evident than increases. Abnormally low levels are more easily concealed by variations due to physiological state such as nutrition (which is difficult to standardize in ill newborns or infants) and to machine performance in diagnostic laboratories. In fact, currently used

diagnostic technologies cannot discriminate low from normal CSF asparagine levels. In summary, results to date suggest that in patients with unexplained congenital encephalopathy with microcephaly, the absence of a low value does not exclude ASNS deficiency. In the future, an enzyme assay may play an important role in the complete diagnostic evaluation of patients suspected of ASNS deficiency but experience is too limited to conclude. In children with severe congenital encephalopathy and microcephaly, ASNS deficiency should be considered, and molecular diagnosis is the only method with proven reliability.

All three known deficiencies of amino acid biosynthesis present mainly with neurological features. In these conditions, the deficient amino acid becomes essential. Hence, an obvious first consideration for therapy is dietary supplementation, to provide the deficient amino acid to the brain. Plasma levels can usually be substantially increased by dietary supplementation and despite the complex transport systems for amino acids at the brain endothelium, a therapeutic benefit of supplementation has been reported in serine biosynthetic disorders and glutamine synthetase deficiency [116,117]. Supplementation with asparagine therefore seems reasonable in ASNS deficiency. However, the prenatal onset of the microcephaly and the early postnatal presentation raise the possibility that such treatment will not be curative unless started prenatally.

The *Asns* mouse we have analyzed here will provide a model for future comprehensive exploration of the factors influencing phenotypic severity. Comparing

this hypomorphic mouse model with a null mouse model will allow us to directly evaluate how residual levels of ASNS activity compare with the absence of ASNS activity, which may inform us about differences in clinical presentation. We can also utilize both animal models when testing the effects of dietary supplementation, which would ensure that a range of ASNS activities were represented, thus covering the full range of ASNS activities that may also occur in patients. This work therefore sets the stage for evaluation of treatment options in *Asns* mouse models.

Early diagnosis of ASNS deficiency is now possible. Careful clinical observations and studies of Asns-deficient mice will help define the clinical spectrum and resolve central unanswered issues regarding the pathophysiology of this condition.

3. Diagnostic exome sequencing in 62 patients with undiagnosed conditions[1]

3.1 Introduction

The molecular basis of more than 3,000 Mendelian disorders is now established. Recently, causal disease genes have frequently been identified through next generation sequencing which has proven very effective in identifying genes that were refractory to linkage analyses [118]. Despite this progress, patients routinely present in the clinic either with unknown conditions or apparently known ones that do not resolve upon traditional genetic testing. In these situations, next generation sequencing is becoming increasingly common, with "success" rates ranging from 16% to about 50% [119-123]. While there is no consensus about appropriate methods for interpreting the genetic data, either directly or indirectly, these studies primarily focused on genes that were already known to cause Mendelian diseases. For example, in the largest study to date Yang *et al.* (2013) studied 250 patients and focused on mutations in genes that were previously shown to cause a disease with a "similar presentation" [122].

We adopted a more comprehensive approach to diagnostic exome sequencing in our cohort of 62 undiagnosed patients, expanding on a framework of identifying

---

[1] This work is part of a currently unpublished collaboration. My collaborators include: Pingxing Xie, Slave Petrovski, Xiaolin Zhu, Yi-Fan Lu, Bruria Ben-Zeev, Andreea Nissenkorn, Yair Anikster, Danit Oz-Levi, Ryan S. Dhindsa, Yuki Hitomi, Kelly Schoch, Rebecca Crimian, Gali Heimer, Dina Marek-Yagel, Yujun Han, Gordon Worley, Jennifer Goldstein, Yong-Hui Jiang, Doron Lancet, Elon Pras, Anna C Need, Vandana Shashi, and David B Goldstein.

qualifying variants throughout the genome; as was first proposed by Need *et al.* 2012 [121]. For each trio, we first identified all qualifying variants: annotated as putatively functional and observed as a novel *genotype* in the probands (not observed in the unaffected parents or in large control datasets). Then, each gene carrying a qualifying variant(s) was analyzed in public databases to determine whether mutations in the gene have been reported to cause similar clinical presentations. We then attempted to extend the discovery paradigm to identify novel genes by using two novel analytic approaches: (1) a joint gene- and variant-level framework for prioritizing *de novo* mutations and (2) protein-protein interaction (PPI) networks seeded by genes carrying qualifying variants.

The basic approach of beginning with a set of qualifying variants, therefore, opens up a new set of analyses – that are independent of established genotype-phenotype relationships – to apply in the interpretation of personal genomes. While this approach runs the risk of false positive genetic diagnoses, it can also lead to new discoveries that would otherwise be missed. For example, Need *et al.* [121] recognized a patient with compound heterozygous truncating variants in *NGLY1,* encoding N-glycanase 1, an enzyme involved in the degradation of misfolded glycoproteins. Despite *NGLY1* having no association with a specific disorder at that time, the phenotype of this child was recognized as being consistent with a congenital disorder of glycosylation. Recently, careful clinical phenotyping has identified eight other unrelated patients with

a homozygous truncating *NGLY1* mutation, thus confirming the identification of a new disorder (OMIM#615273, manuscript in preparation).

From a biological perspective, understanding the etiology of additional rare neurodevelopmental disorders informs our understanding of the roles of these genes in healthy individuals and the clinical spectrum associated with their dysfunction. From a clinical perspective, there is great value to obtaining a genetic diagnosis. Attaining a genetic diagnosis improves prognosis counseling, facilitates discussion of recurrence risk, and may impact medical management (drug therapy) in some cases.

The approaches introduced above have been used to analyze exome sequence data from 62 patients with a diverse range of undiagnosed or unresolved disorders (Table 13) and their unaffected biological parents (trios). The vast majority of these patients (85%) suffer from neurodevelopmental disorders.

3.2 Materials and methods

3.2.1 Recruitment of subjects and collection of samples

A total of 62 patients with severe undiagnosed or unresolved rare genetic disorders, and their unaffected biological parents, participated in this study (Figure 21). Twenty-four of the trios were recruited at the Genome Sequencing Clinic at Duke University Medical Center. Another 32 were recruited at the pediatric clinic of the Sheba Medical Center in Tel Hashomer, Israel. We previously published an analysis of 12 trios

from Duke University Medical Center [121], and here we have also fully re-analyzed the six trios from that study that were not previously resolved.

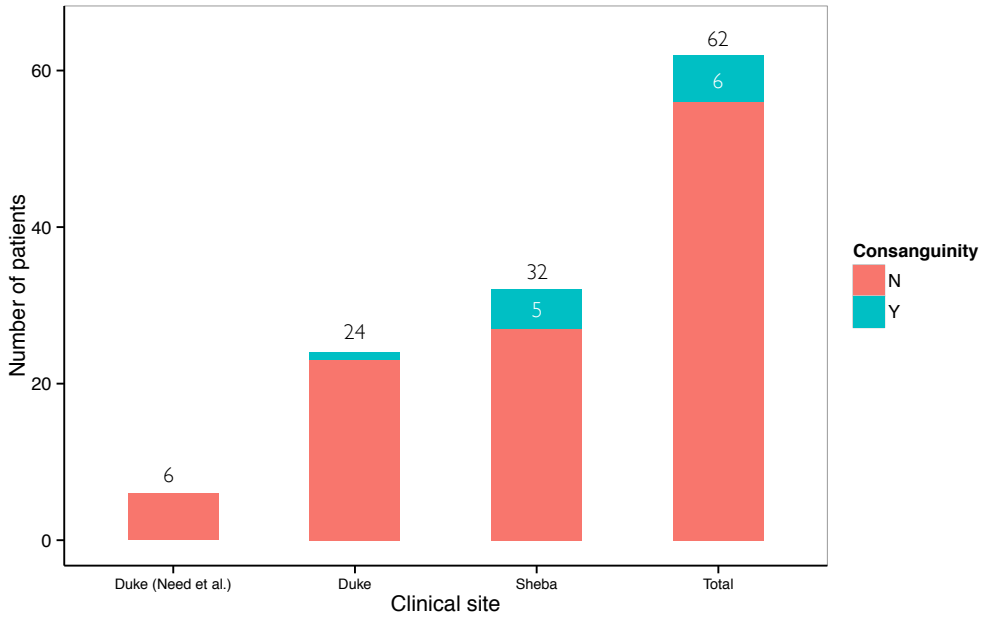


Figure 21. Breakdown of 62 patients by clinical site with patients from consanguineous unions labeled.

Patients recruited for this study had diverse clinical phenotypes and all patients were suspected to have an underlying genetic diagnosis (e.g., no evidence of birth asphyxia or non-accidental trauma). Despite the diversity of clinical phenotypes (Table 13), 85% of cases have a neurodevelopmental disorder; this is based on a broad definition of observing any of the following phenotypic features: ID, developmental delays, language delays, microcephaly or other brain malformations, or seizure disorders.

As is common in traditional genetic diagnostics, most of the 62 trios were screened in the clinic for any individual genes with a high probability of causing the observed disorder when mutated (e.g., *SCN1A* mutations in epileptic encephalopathy patients), as well as tested for chromosomal abnormalities and copy number variations. At the time of enrollment, however, all patients remained genetically unresolved. To be eligible for this study, the proband must have undergone a clinical genetics evaluation and not received a diagnosis after completion of any pertinent molecular testing. Thus the attending physician excluded all readily identifiable genetic conditions and non-genetic causes of their conditions. The appropriate Institutional Review Boards approved this research protocol. Written informed consent was received from all participants or their guardians.

We used two sources of control data in this study. The primary control cohort consisted of subjects enrolled in Center for Human Genome Variation studies through Duke Institutional Review Board approved protocols (n=1300). The secondary control cohort were the subjects enrolled in the NHLBI GO Exome Sequencing Project (ESP, n=6503).

Table 13. Phenotypic descriptions for undiagnosed patient cohort.

Trio numbers starting with "N" were trios previously sequenced and unresolved when originally reported(need et al). ND disorder = suspected neurodevelopmental disorder.

Trio	Consan-guinity	Recruitment Site	Sex	Clinical Phenotype	ND disorder
1	Y	Sheba	M	Profound intellectual disability, Angelman features, early epileptic encephalopathy.	Y
2	Y	Sheba	F	Melkersson-Rosenthal- like syndrome. Facial swelling (beginning on the lips and later involving the face). Morbid obesity and metabolic syndrome, abnormal skin striae, and depression.	Y
3	N	Sheba	M	Severe intellectual disability, myoclonic seizures (generalized polyspike and wave), mild brain atrophy, hemolytic anemia with hyperactive bone marrow.	Y
4	N	Sheba	M	Pitt-Hopkins features, microcephaly, hyperventilation, failure to thrive.	Y
5	N	Sheba	M	Mild global delay, dementia, Parkinsonism.	Y
6	N	Sheba	M	Progressive cerebelocerebllar atropy with profound intellectual disability, epilepsy, cortical blindness, peripheral edema, and is suspected of having pontocerebellar hypoplasia.	Y
7	N	Sheba	F	Epileptic encephalopathy, myoclonic jerks from day 1, profound intellectual disability, cardiomyopathy, cortical blindness, tracheostomy, PEG feeding, hydronephrosis, progressive cerebellar and cerebral atrophy, and leukoencephalopathy.	Y
8	N	Sheba	M	Leigh syndrome, regression at 7-8 month of age with hypsarrhythmia, basal ganglia and brain stem necrosis and cortical atrophy.	Y
9	N	Sheba	F	Leukoencephalopathy, mild developmental delay, slight delay in language development, mild ataxia, epilepsy, and hypergammaglobulinemia.	Y
10	N	Sheba	M	Epileptic encephalopathy with severe intellectual disability and macrocephaly.	Y
11	N	Sheba	F	Severe developmental delay, hypotonia, ptosis, hyporeflexia, high creatine kinase, cerebella malformation, colpocephaly.	Y
12	N	Sheba	M	Developmental delay, profound intellectual disability, seizures, apnea, hypotonia, hypoxic ischemic encephalopathy, micropenis, pontocerebellar hypoplasia, cerebral atrophy and thin corpus callosum.	Y
13	N	Sheba	M	Adult onset rhabdomyolysis after physical activity	N
14	N	Sheba	M	Carnitine deficiency, familial renal fancony with subsequent carnitine deficiency, abnormal motor control, short stature, polyuria, and convulsive disorder.	N
15	N	Sheba	F	FTT, autonomic neuropathy, ID, brain atrophy with white matter changes.	Y
16	N	Sheba	M	Suspected O-Linked glycosylation disorder, developmental delay, conduct disorder, mild myopathy, speech disorder, tantrums, abdominal hypopigmentary lesions, and resolved hand tremor.	Y
17	N	Sheba	M	Hallermann-Streiff syndrome, CHF due to restrictive cardiomyopathy, chronic lung disease and pulmonary hypertension.	N
18	Y	Sheba	F	Leigh syndrome.	Y

Trio	Consan-guinity	Recruitment Site	Sex	Clinical Phenotype	ND disorder
19	N	Sheba	M	Severe hypotonia, gastroesophageal reflux diseasareflexia, intellectual disability, autonomic dysfunction with encephalopathic events, breathing abnormalities.	Y
20	N	Sheba	F	Microcephaly, brain atrophy.	Y
21	N	Sheba	M	Congenital adrenal hyperplasia-like, obese, early pubarche, advanced bone age, acne, high cortisol in the urine, elevated androstenedione, elevated Igf1.	N
22	N	Sheba	M	Normocephalic leukoencephalopathy with subcortical cysts.	Y
23	N	Sheba	F	Hypsarrhythmia from 5 months age, profound MR, simplified gyral pattern microcephaly.	Y
24	N	Sheba	M	Leigh-like features, pallidum necrosis.	Y
25	N	Sheba	F	Rett-like feature, global delay, microcephaly.	Y
26	N	Sheba	F	Basal ganglia necrosis.	Y
27	Y	Sheba	M	Patient was healthy until 3.5 years then developed epilepsy with drop attacks, blindness (retinitis pigmentosa).	Y
28	Y	Sheba	M	Severe global delay, spastic quadriparesis, pontocerebellar hypoplasia, cataract, PCH.	Y
29	N	Sheba	F	Hereditary spastic paraparesis, amyotrophy, intellectual disability, behavior problems, short stature, microcephaly.	Y
30	N	Sheba	F	Ketotic hypoglycemia.	N
31	N	Sheba	F	Cockayne syndrome-like features.	Y
32	N	Sheba	M	Early-onset strokes and epilepsy.	Y
33	N	Duke	F	Intellectual disability, microcephaly, mild to moderate hypotonia, hypermotoric.	Y
34	N	Duke	M	Elevated liver transaminases along with minor dysmorphic features, minor learning difficulties, no developmental delay.	N
35	N	Duke	M	Overgrowth, dysmorphic features, mild developmental delays, and macrocephaly.	Y
36	N	Duke	M	Global developmental delays with no speech, gastroschisis, neonatal sigmoid venous sinus thrombosis, acqueductal stenosis causing hydrocephalus and shunt placement, abnormal optic nerves, hypopituitarism, myopia, femoral retroversion, behavior problems, dysmorphic.	Y
37	N	Duke	F	Hypotonia, global developmental delays (does not sit or speak), microcephaly with progressive cerebellar and vermian atrophy.	Y
38	N	Duke	F	Seizures and developmental delays, with regression with onset of seizures, abnormal citrulline on amino acid.	Y
39	N	Duke	F	Aqueductal obstruction with hydrocephalus, infantile spasms, other midline brain abnormalities, vertebral abnormalities, and bilateral retinal abnormalities.	Y

Trio	Consan-guinity	Recruitment Site	Sex	Clinical Phenotype	ND disorder
40	N	Duke	M	Profound developmental delays with failure to thrive, craniosynostosis, cortical visual impairment, glaucoma, supravalvular pulmonic stenosis.	Y
41	N	Duke	F	Microcephaly, developmental delay and hypotonia, seizures.	Y
42	N	Duke	M	Severe intractable seizures with developmental delay.	Y
43	N	Duke	M	Global developmental delays, oculomotor apraxia, elevated alpha-fetoprotein, and lack of tears.	Y
44	N	Duke	F	Right-sided choanal atresia, bilateral sensorineural hearing loss, right sided ptosis, mild optic nerve hypoplasia, a Klippel-Feil anomaly, a history of an ASD with pulmonary stenosis, and developmental delays.	Y
45	N	Duke	F	Peter's anomaly bilaterally with microphthalmia, more on the right than the left, bilateral choanal atresia, communicating hydrocephalus, bilateral hearing loss requiring cochlear implants and acute lymphoblastic leukemia that has been in remission.	N
46	N	Duke	M	Profound hypotonia, severe developmental delays, involuntary movements that are thought to be choreoathetotic in nature as well as frequent arching of his back that is thought to be related to seizure activity.	Y
47	N	Duke	F	Severe developmental delay, hypotonia, microcephaly, esotropia, growth failure, static encephalopathy, seizures and scoliosis	Y
48	N	Duke	M	Progressive developmental decline, pontocerebellar atrophy and diffuse mild cerebral atrophy.	Y
49	N	Duke	M	Cardiomegaly, pulmonary hypertension, Von Willebrand disease and congenital heart anomalies.	N
50	N	Duke	F	Microcephaly, minor dysmorphisms, profound developmental delays, epilepsy, gingival hypertrophy, microdontia, intracranial calcification of the basal ganglia, white matter abnormalities.	Y
51	Y	Duke	F	Autism spectrum disorder, developmental delay, and osteopetrosis.	Y
52	N	Duke	F	Myoclonic seizures, absence seizures, generalized tonic clonic seizures, developmental regression.	Y
53	N	Duke	F	Hypertonia, opisthotonous movement, seizure, and developmental regression.	Y
54	N	Duke	M	Atrial septal defect, intrauterine growth restriction, microcephaly, sensory integration d/o and dyspraxia, developmental delay, strabismus, communicating hydrocele.	Y
55	N	Duke	M	Athetoid cerebral palsy, developmental delays, seizure, recurrent vomiting.	Y
56	N	Duke	F	Progressive epileptic encephalopathy with neurological decline beginning around 2 1/2 to 3 years of age.	Y
N4	N	Duke	F	Multiple congenital abnormalities and macular degeneration	N
N6	N	Duke	M	Intellectual disability, epilepsy, panhypopituitarism, hypertension, bifid great toe, vertebral segmentation anomalies and sagittal cleft of the vertebra, hypoplastic 13th rib, and delayed bone age.	Y
N8	N	Duke	M	Bicuspid aortic valve, bilateral coronal craniosynostoses, dysmorphic features evident, quadriplegic cerebral palsy, bilateral inguinal hernias, G-tube placement, obstructive sleep apnea, and severe intellectual disability.	Y

Trio	Consan-guinity	Recruitment Site	Sex	Clinical Phenotype	ND disorder
N9	N	Duke	F	Developmental delay, bilateral congenital cataracts and strabismus, ventricular and atrial septal defects, a unilateral clubfoot, and unilateral choanal atresia.	Y
N10	N	Duke	M	Attention deficit hyperactivity disorder, language delays, coarse facial features, bilateral mandibular cysts, low muscle tone.	Y
N12	N	Duke	F	Failure to thrive, borderline microcephaly, dysplastic nails, ventricular septal defect, hip dysplasia, and speech delay.	Y

3.2.2 Exome sequencing

DNA was extracted from a whole blood sample from each participant. To capture the protein-coding and flanking intronic-exonic regions of the genome, we used Illumina TruSeq Exome Enrichment 65Mb Kit (Illumina, San Diego, CA) for 36 trios, Roche NimbleGen SeqCap EZ Exome library 64Mb Kit (Roche NimbleGen, Madison, WI) for 14 trios, and Agilent SureSelect Human All Exon 50Mb Kit (Agilent, Santa Clara, CA) for 12 trios. The capture kit was always consistent within all members of a given trio. All sequencing was performed on the Illumina HiSeq2000 platform (Illumina, San Diego, CA) in the Genomic Analysis Facility in the Center for Human Genome Variation at Duke University.

After sequencing, PCR duplicates were removed using Picard software and reads were aligned to Genome Reference Consortium Human Genome build 37 (GRCh37) using the Burrows-Wheeler Alignment Tool [83], and variants were called using the Genome Analysis Toolkit (GATK) [86] and annotated using SnpEff (http://snpeff.sourceforge.net/). The six re-analyzed trios [121] were also subjected to this bioinformatics pipeline; previously, they were aligned to an older reference genome (GRCh36) and variants were called using SAMtools [84]. The 62 trio samples had an average coverage of >78x and on average >94% of targeted regions had a coverage of ≥5.

3.2.3 Identification of qualifying variants

We focused on variants that were annotated as functional: nonsynonymous, truncating, variants in essential splice sites, or small insertion/deletion variants (indels) located in the protein-coding regions. For all such variants, we screened for those conforming to one of four genetic models (Figure 22): (1) *de novo* variants in the proband (including X chromosome *de novo* variants) which were not observed in the parents nor any of the controls; (2) homozygous variants in the proband which were heterozygous in both parents, never homozygous in the controls, and had minor allele frequency (MAF) < 1% in controls; (3) hemizygous X-chromosome variants in male probands that were inherited from the unaffected mother, never observed in any male controls, and had MAF < 2% in female controls; (4) compound heterozygous variants in the proband (one variant inherited from each parent), where each of the two compound heterozygous variants were never homozygous in the controls, and individually had MAF < 1% in the controls. Additionally, the frequency of co-occurrence of the two contributing variants had to be zero in the CHGV control cohort. Variants meeting these criteria are referred to as "qualifying variants" and the genes harboring these variants are referred to as "qualifying genes".

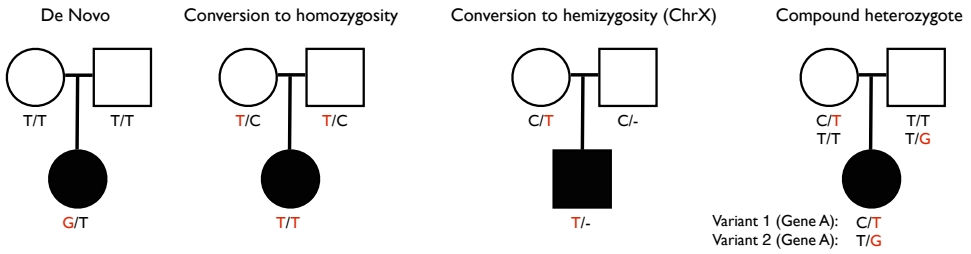


Figure 22. Schematic of inheritance patterns for qualifying variants of interest in a proband with an undiagnosed disorder.

All qualifying variants passed the following quality filters: (1) read depth ≥ 10 in all three family members of a trio; (2) for *de novo* variants, <5% of the reads in either parent should carry the alternate allele; (3) for heterozygous variants, ≥25% of the reads should carry the alternate allele; (4) visual inspection using Integrative Genomics Viewer (IGV) confirmed that the variant was not called due to alignment errors.

3.2.4 Analysis of qualifying variants

Sanger sequencing and/or CLIA testing were performed to validate the candidate causal mutation(s) with the proband of each trio (Table 14). For the Israeli patients, all available family members were also checked for cosegregation. We did not comprehensively validate all qualifying variants by Sanger sequencing.

3.2.4.1 Qualifying variants in genes already implicated in similar phenotypes

For each qualifying gene, we considered whether it had been implicated in a closely matching phenotype by searching the Online Mendelian Inheritance in Man

(OMIM) database and PubMed. When genes were implicated, we then asked whether the precise mutations were reported in the literature to cause a closely matching phenotype, and if not, whether the mutations were evidently within the same class as already known disease causing mutations (e.g., dominant, recessive, truncation).

While the screen for qualifying variants required the absence of the relevant genotypes in both control sets, there were two exceptions to this rule. In one case (family 19, *TECPR2*), the relevant genotype in the NHLBI ESP database was found in a sample with an average read depth of only 17 and thus this variant was judged to be likely just a heterozygote genotype mistakenly called as a homozygote due to low coverage (confirmed by visual inspection of alignments at this site via personal communication with NHLBI ESP project). In the second case (family 47, *ASXL1*), a well-documented disease-causing variant was observed in two NHLBI ESP subjects (of African ancestry). Because this *ASXL1* variant (p.R404*) did not meet our qualifying variant criteria, we considered this mutation to be only a good candidate in family 47, rather than declaring this as the genetic diagnosis. However, we note that no systematic screen for such exceptions was undertaken, unless the exact variant or variant class (e.g., frameshift) was previously listed as disease-causing.

3.2.4.2 Joint gene- and variant-level prioritization of *de novo* mutations

For all trios that were not resolved by an inherited genetic model we used a gene- and variant-level prioritization framework to assess *de novo* mutations, similar to what was presented by Petrovski *et al*. [124].

Filtering for high confidence *de novo* mutations

This systematic approach requires high quality variant calls. Therefore, we used a strict approach to identify and filter the *de novo* mutations identified in our 62 trios. Putative *de novo* mutations were jointly called with the GATK Unified Genotyper for all family members in a trio. For all putative *de novo* mutations, we removed variants not meeting the following criteria: (1) the read depth in both parents ≥10; (2) the depth of coverage in the proband should be ≥0.1% of the sum of the coverage in both parents; (3) for de novo variants, <5% of the reads in either parent carries the alternate allele; (4) ≥25% of the reads in the proband support the alternate allele; (5) the normalized, phred-scaled likelihood (PL) scores for the proband genotypes AA, AB, and BB, where A is the reference allele and B is the alternate allele, should be >20, 0, and >0, respectively; (6) the PL scores for both parents should be 0, >20, and >0; (7) ≥three variant alleles must be observed in the proband; and (8) the *de novo* mutation had to be located within CCDS Release 9 genic boundaries, with 2 base pair extensions on either side of protein-coding exons. PL scores are assigned such that the most likely genotype is given a score of 0, and the

score for the other two genotypes represent the likelihood that they are not the true genotypes. We also used published *de novo* mutation calls from 610 published control trios; these are neurologically healthy siblings of autism spectrum disorder (ASD), ID, and epileptic encephalopathy patients [62,63,125,126].

Gene-level score

For the gene-level score, we used the Residual Variation Intolerance Score (RVIS) introduced by Petrovski *et al*. [124].

Variant-level score

High quality *de novo* mutations from the 62 case trios and 610 published control trios were annotated using Ensembl Variant Effect Predictor (Ve!P version 2.6). PolyPhen-2 scores for missense *de novo* mutations were taken from Ve!P annotations; when multiple CCDS transcripts of a gene were affected by the mutation, the most damaging CCDS PolyPhen-2 score was used. When PolyPhen-2 scores were classified as "unknown" by Ve!P, we adopted the PolyPhen-2 scores using the PolyPhen-2 web interface "CCDS" transcript option. In contrast to Petrovski *et al.* [124], we incorporated nonsense, essential splice site, and indel coding *de novo* mutations by assigning them a PolyPhen-2 probabilistically damaging score of 1, and assigning synonymous *de novo* mutations a PolyPhen-2 probabilistically damaging score of 0.

Exclusions

For individuals with multiple *de novo* mutations, we considered only the most damaging *de novo* mutation, assessed by Euclidean distance from the most damaging coordinate (i.e., PolyPhen-2 score of 1 [x-axis], and RVIS percentile score of 0 [y-axis]). This differed from Petrovski *et al.* [124] where the most damaging *de novo* mutation was selected based on the lowest RVIS.

We excluded *de novo* mutations in genes without an available RVIS (n=12), and also excluded those present in the NHLBI ESP database (n=46). A total of 58 *de novo* mutations observed in the control trios were excluded, while no *de novo* mutations in our patients were excluded during these two steps.

3.2.4.3 PPI networks seeded by genes carrying qualifying variants

For patients who were not resolved by identification of a qualifying variants already implicated in a closely matching phenotype, we searched for genes that are directly connected in PPI networks to those carrying qualifying variants using the NCBI Interactions database. For every patient, we then checked these neighboring genes in the OMIM database to see if any of them have been implicated in similar phenotypes.

3.3 Results

3.3.1 Qualifying variants in genes with established clinical relevance

We analyzed 62 exome sequenced trios and made a comprehensive list of all qualifying variants (Figure 22) in each genome. On average, 13 genes per patient carried qualifying variants. All genes carrying qualifying variants were examined to determine if any of those genes were already implicated in diseases with similar phenotypes. By protocol, a case was considered to have obtained a genetic diagnosis if the treating clinician and the research genetics team agreed that a genetic diagnosis was achieved; this decision was based on observing a qualifying variant(s) in a gene that was associated with a closely matching condition, where either the precise mutations had been seen before or the same class of mutation was previously established as disease-causing for the studied disorder (e.g., recessive, truncating). However, we recognize that some small fraction of these resolved cases may in fact be incorrect. To be conservative, if the treating clinician and research genetics team considered a variant as potentially linked to the patient's phenotype, we referred to these cases as having a good candidate variant.

We did not proactively look for incidental or secondary findings in genes that were unrelated to the proband's immediate phenotype; however, if we encountered a risk for a disease known to potentially cause premature death if untreated, it would be reported back to the families. We did not identify any such findings.

This approach resulted in a genetic diagnosis in 18 (32%) of the 56 cases, and two (33%) of the six previously unresolved cases [121] (Figure 22, Figure 23, Table 14). Another 11 patients, plus one of the six previously unresolved cases, had good candidate variants (Figure 22, Figure 23, Table 14).

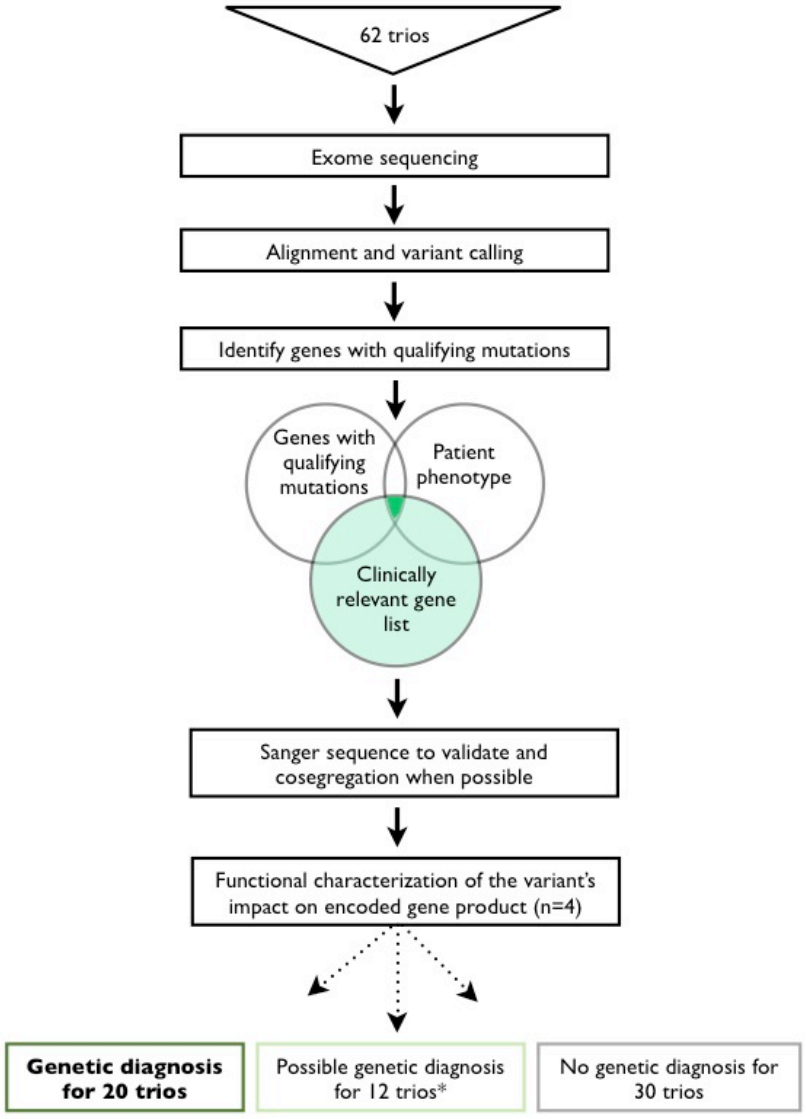


Figure 23. Flow chart of the analysis of 56 newly recruited trios and 6 previous unresolved trios based on the established clinical relevance of the genes.

Of the four considered modes of inheritance, each was found to be disease-causing in at least once instance; including: four hemizgyotes, five compound heterozygotes, five homozygotes and six *de novo* mutations (Figure 24).

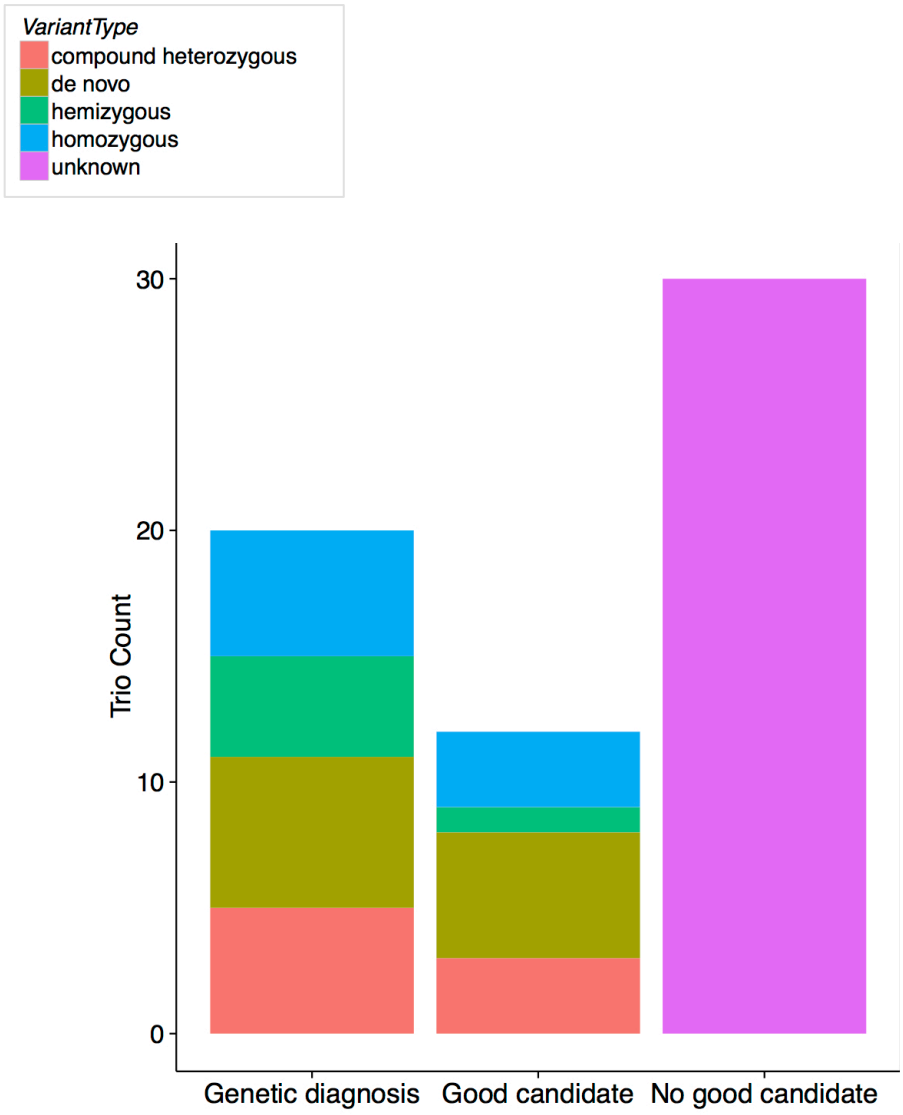


Figure 24. Proportion of disease-causing and candidate disease-causing mutations by mode of inheritance.

Table 14. Qualifying mutation details for trios where a genetic diagnosis or candidate disease-causing variant was identified.

Trio[a]	Diagnosis status[b]	Candidate genes	Variant type	Function	Transcript change	Amino acid change	MAF in internal controls	MAF in NHLBI ESP	Associated disease (OMIM#) or protein function
1	D	*PIGA*	hemizygous	missense	NM_002641.3:c.1352T>C	NP_002632.1:p.Ile451Thr	0	0	Multiple congenital anomalies-hypotonia-seizures syndrome (300868)
2	C	*SLC17A1*	homozygous	nonsense	NM_005074.3:c.1287_1292del	NP_005065.2:p.Trp429*	0	0	Sodium-dependent phosphate transporter in kidney
3	D	*GPI*	homozygous	missense	NM_000175.3:c.1066G>A	NP_000166.2:p.Asp356Asn	0	0	Hemolytic anemia (613470)
6	C	*SHQ1*	compound heterozygous	missense; missense	NM_018130.2:c.[1277C>T];[1003C>T]	NP_060600.2:p.[Ala426Val];[Arg335Cys]	0;0	0;0	Mutations in *SHQ1* are not associated with any known disorders, however other genes involved in rRNA processing have been associated with PCH (614678)
7	D	*PIGN*	compound heterozygous	missense; splice site	NM_012327.5:c.[675-1G>C];[1261T>C]	IVS9-1C>G; NP_036459.1:p.[Cys421Arg]	0;0	0;0	Multiple congenital anomalies-hypotonia-seizures syndrome 1 (614080)
10	D	*KCNQ2*	*de novo*	missense	NM_004518.4:c.841G>A	NP_004509.2:p.Gly281Arg	0	0	Epileptic encephalopathy (613720)
15	D	*DST*	compound heterozygous	missense; missense	NM_020388.3:c.[8156G>A];NM_015548.4:c.[5951C>A]	NP_065121.2:p.[Gly2719Glu];NP_056363.2:p.[Pro1984His][c]	0.006; 0.008	0.007; 0.01	Hereditary sensory and autonomic neuropathy (614653)
18	D	*NDUFAF2*	homozygous	missense	NM_174889.4:c.1A>T	NP_777549.1:p.Met1Leu[d]	0	0	Leigh syndrome (256000)
19	D	*TECPR2*	homozygous	frameshift	NM_001172631.1:c.1318_1319del	NP_001166102.1:p.Leu440Valfs*13	0.0004	0.0006	Spastic paraplegia 49, autosomal recessive (615031)

Trio[a]	Diagnosis status[b]	Candidate genes	Variant type	Function	Transcript change	Amino acid change	MAF in internal controls	MAF in NHLBI ESP	Associated disease (OMIM#) or protein function
22	C	*SUOX*	homozygous	missense	NM_000456.2:c.739C>A	NP_000447.2:p.Leu247Met	0	0	Sulfite oxidase deficiency (272300)
24	D	*PDHA1*	hemizygous	missense	NM_000284.3:c.214C>T	NP_000275.1:p.Arg72Cys [e]	0	0	Leigh syndrome, X-linked (308930)
26	D	*GCDH*	compound heterozygous	missense; missense	NM_000159.2:c.[301G>A]; [1213A>G]	NP_000150.1:p.Gly101Arg; [Met405Val] [f]	0.0004;0	0;0.0003	Glutaricaciduria, type I (231670)
27	D	*CLN6*	homozygous	missense	NM_017882.2:c.443T>A	NP_060352.1:p.Val148Asp	0	0	Ceroid lipofuscinosis, neuronal, 6 (601780)
28	C	*MED29*	homozygous	missense	NM_017592.1:c.416T>C	NP_060062.1:p.Leu139Pro	0	0	Microcephaly, postnatal progressive, with seizures and brain atrophy (613668)
33	D	*ASXL3* [c]	*de novo*	frameshift	NM_030632.1:c.4151_ 4153del	NP_085135.1:p.Thr1384_Val 1385delinsIle	0	0	Bohring-Opitz like syndrome (NA)
34	D	*GNE* [c]	*de novo*	missense	NM_001128227.2:c.890G>A	NP_001121699.1:p.Arg297Gln[g]	0	0	Sialuria (269921)
35	C	*ASXL2*	*de novo*	frameshift	NM_018263.4:c.2424_ 2425del	NP_060733.4:p.Thr809Gln fs*17	0	0	Bohring-Opitz like syndrome caused by ASXL1 and ASXL3 (605039 and NA)
36	C	*BCOR*	hemizygous	missense	NM_001123383.1:c.2008C>T	NP_001116855.1:p.Pro670Ser	0	0.0002	Microphthalmia, syndromic 2 (300166)
37	D	*SEPSECS* [c]	compound heterozygous	missense; splice site	NM_016955.3:c.[388+3A>G] ; [1A>G]	IVS3+3C>T; NP_058651.3:p.[Met1Val]	0;0.0008	0;0	Pontocerebellar hypoplasia type 2D (613811)
38	C	*GFAP*	*de novo*	missense	NM_001131019.2:c.1004G>T	NP_001124491.1:p.Gly335Val	0	0	Alexander disease (203450)

Trio[a]	Diagnosis status[b]	Candidate genes	Variant type	Function	Transcript change	Amino acid change	MAF in internal controls	MAF in NHLBI ESP	Associated disease (OMIM#) or protein function
41	C	*ATP8A2*	*de novo*	nonsense	NM_016529.4:c.1110C>A	NP_057613.4:p.Tyr370*	0	0	Cerebellar ataxia, mental retardation, and dysequilibrium syndrome 4 (615268)
42	C	*ANK3*	compound heterozygous	missense; missense	NM_020987.3:c.[4465C>T]; NM_001149.3:c.[2380G>T]	NP_066267.2:p.[Pro1489Ser]; NP_001140.2:p.[Asp794Tyr][i]	0.006;0.002	0.005;0.002	Ankyrin 3
43	C	*RYR1*	compound heterozygous	missense; missense	NM_000540.2:c.[4507G>A]; [10505G>T]	NP_000531.2:p.[Gly1503Arg]; [Arg3502Leu]	0;0	0.0001;0	Minicore myopathy with external ophthalmoplegia (255320)
46	D	*GNAO1*[c]	*de novo*	missense	NM_020988.2:c.124G>C	NP_066268.1:p.Gly42Arg	0	0	Epileptic Encephalopathy[m]
47	C	*ASXL1*[c]	*de novo*	nonsense	NM_015338.5:c.1210C>T	NP_056153.2:p.Arg404*[h]	0	0.0002	Bohring-Opitz syndrome (605039)
51	D	*SNX10*[c]	homozygous	missense	NM_001199835.1:c.284G>A	NP_001186764.1:p.Arg95His	0	0	Osteopetrosis, autosomal recessive 8 (615085)
52	D	*CLN6*[c]	compound heterozygous	missense; insertion	NM_017882.2:c.[296A>G]; [220_221insGGT]	NP_060352.1:p.[Lys99Arg]; [Trp73dup]	0;0	0.0001;0	Ceroid lipofuscinosis, neuronal, 6 (601780)
55	D	*SLC6A8*	hemizygous	inframe deletion	NM_001142805.1:c.321_323del	NP_001136277.1:p.Phe107del	0	0	Cerebral creatine deficiency syndrome 1 (300352)
56	D	*KCNT1*	*de novo*	missense	NM_020822.2:c.2386T>C	NP_065873.2:p.Tyr796His[i]	0	0	Early infantile epileptic encephalopathy (614959)
N6	C	*HNRNPU*	*de novo*	splice site	NM_031844.2:c.1615-1G>A	IVS9-1G>A	0	0	Heterogeneous nuclear ribo-nucleoprotein U, implicated in epileptic encephalopathy

Trio[a]	Diagnosis status[b]	Candidate genes	Variant type	Function	Transcript change	Amino acid change	MAF in internal controls	MAF in NHLBI ESP	Associated disease (OMIM#) or protein function
N6	C	*SMAD1*	*de novo*	missense	NM_001003688.1:c.511A>G	NP_001003688.1:p.Thr171Ala	0	0	Bone development
N8	D	*ATRX*[c]	hemizygous	missense	NM_000489.3:c.736C>T	NP_000480.2:p.Arg246Cys[j]	0	0	Alpha-thalassemia; mental retardation syndrome (301040)
N12	D	*SRCAP*[c]	*de novo*	nonsense	NM_006662.2:c.7330C>T	NP_006653.2:p.Arg2444*[k]	0	0	Floating-Harbor syndrome (136140)

[a] Family numbers starting with "N" were trios previously sequenced and unresolved when originally reported[121]. There are two candidates listed in trio N6, both are related to different aspects of the observed phenotype (the *SMAD1* mutation may be related to the bone abnormalities since this is a known function of this protein).

[b] D = Genetic diagnosis determined; C = good candidate

[c] Sanger sequencing was performed to validate the qualifying variants listed in this table. Genes with an asterisk indicate CLIA testing was performed for validation.

[d] This exact mutation in NDUFAF2 has been previously reported in a syndrome with overlapping phenotypic features[127].

[e] This exact mutation in PDHA1 has been previously reported in a syndrome with overlapping phenotypic features[128].

[f] Both mutations forming this GCDH compound heterozygote have been previously reported (G101R as a homozygote[129]; M405V as a compound heterozygote[130]) in a syndrome with overlapping phenotypic features.

[g] This exact mutation in GNE (reported as R266Q based on a different protein isoform) has been previously reported in a syndrome with overlapping phenotypic features[131].

[h] This exact mutation in ASXL1 has been previously reported in a syndrome with overlapping phenotypic features[132].

[i] This exact mutation in KCNT1 has been previously reported in a syndrome with overlapping phenotypic features[133].

[j] This exact mutation in ATRX has been previously reported in a syndrome with overlapping phenotypic features[134].

[k] This exact mutation in SRCAP has been previously reported in a syndrome with overlapping phenotypic features[135].

[l] Both mutations forming the compound heterozygotes in DST and ANK3 are not coding on the same transcripts/protein isoforms.

[m] Evidence for GNAO1 in epileptic encephalopathies comes from [42,136].

3.3.1.1 De novo qualifying variants

Each patient carried an average of 1.2 qualifying *de novo* mutations in the protein-coding regions, ranging from 0 to 5. Six patients were found to carry disease-causing *de novo* mutations, including missense mutations in *KCNQ2* (p.Gly281Arg; family 10), *GNE* (p.Arg297Gln; family 34), *GNAO1* (p.Gly42Arg; family 46), *KCNT1* (p.Tyr796His; family 56), a frameshift mutation in *ASXL3* (p.Thr1384_Val1385delinsIle; family 33), and a nonsense mutation in *SRCAP* (p.Arg2444*; family N12).

Interestingly, we observed three patients with a *de novo* truncating mutation in one of three closely related genes: *ASXL1* (family 47), *ASXL2* (family 35), and *ASXL3* (family 33). Truncating mutations in *ASXL1* and *ASXL3* have been reported to cause Bohring-Opitz[132] and Bohring-Opitz like syndrome [137], respectively. The presentation of the patient from family 47, with the *ASXL1* mutation, was consistent with Bohring-Opitz syndrome and this disease-causing mutation was previously reported [132]; however, due to the presence of this mutation in controls we could not confidently say this mutation was causal (see methods). The patient from family 33, with the *ASXL3* mutation, has features consistent with previously described patients [137]. The patient with the *ASXL2* truncating mutation has macrocephaly and overgrowth phenotypes, both of which are opposite to the phenotypic features found in the patients carrying mutations in *ASXL1* and *ASXL3*. We hypothesize that the *de novo* truncating

mutation in *ASXL2* is the causal mutation in this patient. However, additional patients with similar phenotypes are required to confirm such a hypothesis.

3.3.1.2 Homozygous qualifying variants

Six patients included in this study were self-reported to be children of consanguineous parents (Figure 21). These patients had an average of 26 qualifying homozygous variants, ranging from 5 to 41, while patients with no reported consanguinity had an average of 1.8 qualifying homozygous variants, ranging from 0 to 9. Homozygous causal mutations were found in three of the six patients from consanguineous marriages, suggesting an increased molecular diagnostic rate (~50%) for this subgroup of patients.

Homozygous causal mutations in these three patients included: missense mutations in *NDUFAF2* (p.Met1Leu; family 18), *CLN6* (p.Val148Asp; family 27), and *SNX10* (p.Arg95His); family 51). All of the previously reported disease-causing mutations in *SNX10* were established as either truncating mutations or mutations that completely destroy protein function [138,139], thus we selected the homozygous mutation in *SNX10* (p.Arg95His) for functional experiments, and found that it caused complete degradation of the protein (data not shown). The primary osteopetrosis phenotype of the patient in family 51 matched previously reported cases carrying truncating mutations in *SNX10* [138,139]. Furthermore, the *SNX10* mutation was also observed in the proband's affected sibling.

Two patients with no reported consanguinity, both from the Israeli Ashkenazi Jewish population, have causal homozygous mutations. The patient from family 3 carried a homozygous missense mutation in *GPI* (p.Asp356Asn), and a patient from family 19 carried a homozygous frameshift mutation in *TECPR2* (p.Leu440Valfs*13). Our group recently discovered a homozygous truncating mutation in *TECPR2* (p.Leu1139Argfs*75) causing autosomal recessive hereditary spastic paraplegia in three Jewish Bukharian families [140]. The unrelated patient reported here has a consistent phenotypic presentation to the previously reported cases and carries a different homozygous truncating mutation also in *TECPR2*.

3.3.1.3 Hemizygous qualifying variants

On average, male patients carried 2.1 qualifying hemizygous mutations, ranging from 0 to 6. Four male patients were found to carry hemizygous causal mutations, including missense mutations in *PIGA* (p.Ile451Thr; family 1), *PDHA1* (p.Arg72Cys) ; family 24), *ATRX* (p.Arg246Cys; family N8), and an in-frame deletion in *SLC6A8* (p.Phe107del; family 55).

Additionally, the patient in family 36 carried a hemizygous missense mutation in *BCOR* (p.Pro670Ser). Mutations in *BCOR* cause Lenz microphthalmia syndrome [141] and although this patient did not exhibit obvious microphthalmia, he had abnormally small optic nerves, which may indicate an atypical case. Researchers in our lab assayed the biological impact of p.Pro670Ser and found that it changed protein abundance for

both tested isoforms (data not shown). While this result gives weak evidence for a downstream biological effect, additional patients with atypical presentation or with this same mutation are needed to clarify the role of this variant.

3.3.1.4 Compound heterozygous qualifying variants

Each patient carried an average of 6.1 qualifying compound heterozygous mutations, ranging from 1 to 13. Five patients were found to carry disease-causing compound heterozygous mutations, including families: 7 (*PIGN*), 15 (*DST*), 26 (*GCDH*), 37 (*SEPSECS*), and 52 (*CLN6*). The patient from family 37 carried a compound heterozygote in *SEPSECS* where one variant causes loss of the start codon (p.Met1Val), and the other variant is located 3-bp downstream of exon 3 (NM_016955.3:c.388+3A>G; IVS3+3C>T). To investigate the molecular impact of these variants, researchers in our lab used a minigene assay and showed that this intronic mutation induced exon 3 skipping (data not shown) leading to a frameshift and premature termination of the protein after 41 amino acids. Both variants resulted in complete degradation of the protein (data not shown). Furthermore, these *SEPSECS* variants were also observed in the affected brother. *SEPSECS* mutations cause autosomal-recessive progressive cerebellocerebral atrophy and profound intellectual disability [142]; closely matching the phenotype of our patient and her brother.

3.3.2 A gene- and variant-level framework for the prioritization of *de novo* mutations

The Residual Variation Intolerance Score (RVIS) provides a measure of how well a gene can "tolerate" the presence of putatively damaging functional variants [124]. Assessment of RVIS scores demonstrated that genes causing Mendelian disorders are significantly more intolerant to functional variation [124]. Given that 85% of patients in this cohort have neurodevelopmental disorders, it is important to state that the established neurodevelopmental disease genes are less tolerant to functional variation and thus the RVIS scores tend to be lower (Figure 25). It was previously shown that combining genic "intolerance" to functional variation (RVIS) along with a PolyPhen-2 scores can facilitate the identification of pathogenic mutations [124]. In particular, *de novo* mutations observed in patients ascertained for neurodevelopmental disorders (epileptic encephalopathies, severe intellectual disability, and autism) showed enrichment in a "hot-zone" defined by variants that are the most likely to impact protein function (as defined by PolyPhen-2 probabilistic score) falling in the most intolerant genes in the genome (lowest RVIS scores). The "hot-zone" falls within the two-dimensional space defined by RVIS percentile ≤0.25 (Figure 25) and PolyPhen-2 score ≥ 0.95, as compared to control trios [124].

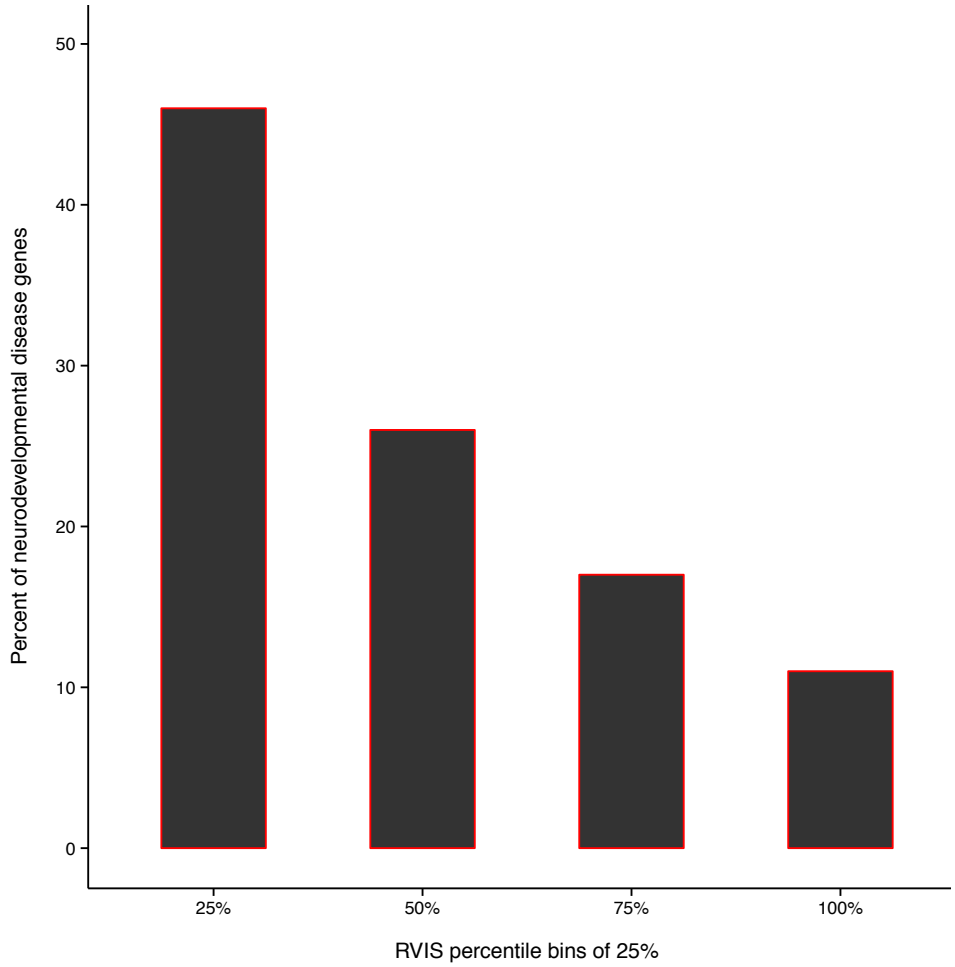


Figure 25. Percentage of neurodevelopmental genes by RVIS percentile. List of neurodevelopmental disease genes from Goh *et al.* [143]. Figure adapted from Petrovski *et al.* [124].

To assess the possible enrichment of *de novo* mutations in the "hot-zone" for our heterogeneous collection of undiagnosed patients, we first obtained high confidence *de novo* mutation calls. This resulted in an average of 0.82 protein-coding *de novo* mutations per proband (Figure 26), which is consistent with previously reported trio sequencing studies [42,58,63,123,125].

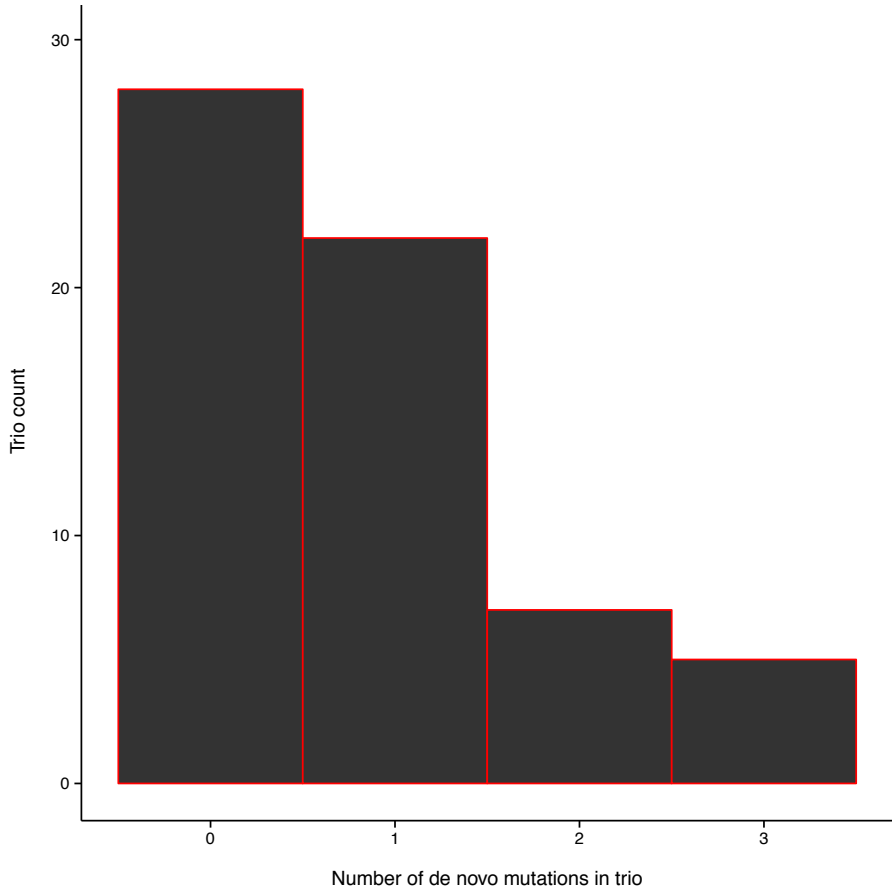


Figure 26. Number of high confidence de novo mutations per trio in all 62 trios.

Next, we determined the expected two-dimensional distribution of *de novo* mutations from previously reported control trios and our undiagnosed patients. Only trios carrying at least one high confidence *de novo* mutation call were included in these analyses. We found that 290 of the 610 controls [62,63,125,126] had at least one *de novo* mutation reported. Taking the single most damaging *de novo* mutation per control revealed that 37 (12.8%) of these *de novo* mutations reside in the hot-zone (Figure 27).

When combining the 56 newly recruited trios with the 12 previously published trios [121] from our group (n=68), we identified 33 trios (23 newly recruited trios and 10 previously published trios [121]) where the disorders were not explained by inherited variants and at least one putative *de novo* mutation was identified. Taking the single most damaging *de novo* mutation per case revealed that 13 *de novo* mutations (39.4%) resided in the hot-zone ($p = 3.5 \times 10^{-4}$, two-tail Fisher's exact test) (Figure 27). Given that the expected proportion of *de novo* mutations in the hot-zone, based on control expectations, is equivalent to 4 out of 33 *de novo* mutations, we observed an excess of nine (69%) hot-zone *de novo* mutations among the undiagnosed patients.

When focusing on only 23 newly recruited trios and five previously unresolved trios who have phenotypes that were not explained by an inherited disease-causing variant (Table 13, Table 14), nine (32.1%) *de novo* mutations were located in the hot-zone (Figure 27, Table 15). The gene-level and variant-level scores used here are, by construction, independent of reported genotype-phenotype associations. Yet, among the nine *de novo* mutations residing in the hot-zone, five are defined as causal mutations based on our assessment of qualifying variants in genes already implicated in similar phenotypes (*KCNQ2* in family 10, *ASXL3* in family 33, *GNAO1* in family 46, *KCNT1* in family 56 and *SRCAP* in family N12), and a sixth one was likely to be causal (*ATP8A2* in family 41) (Table 14). In addition, we recently identified *de novo* truncating mutations in *HNRNPU* and *HNRNPH1* in epileptic encephalopathy patients [42], highlighting a

possible role for these RNA binding proteins in epilepsy phenotypes. The proband from family N6 suffers from epilepsy, among other neurological phenotypes and bone abnormalities (Table 13), and harbors a hot-zone *de novo* mutation in the intronic region of *HNRNPU*. To test the biological impact of this intronic mutation, researchers in our lab used a minigene assay and found that this mutation (NM_031844.2:c.1615-1G>A; IVS9-1G>A) leads to a single nucleotide shift in the intron-exon junction and, consequently, a pre-mature stop codon in exon 9. Semi-quantitative PCR revealed a dramatic reduction in mRNA levels, potentially due to nonsense mediated decay (data not shown). This finding suggests that this mutation may alter the normal activity levels for this RNA binding protein. We highlight this mutation since it supports a possible role HNRNPU in epilepsy; however the current evidence is not strong enough to declare a genetic diagnosis.

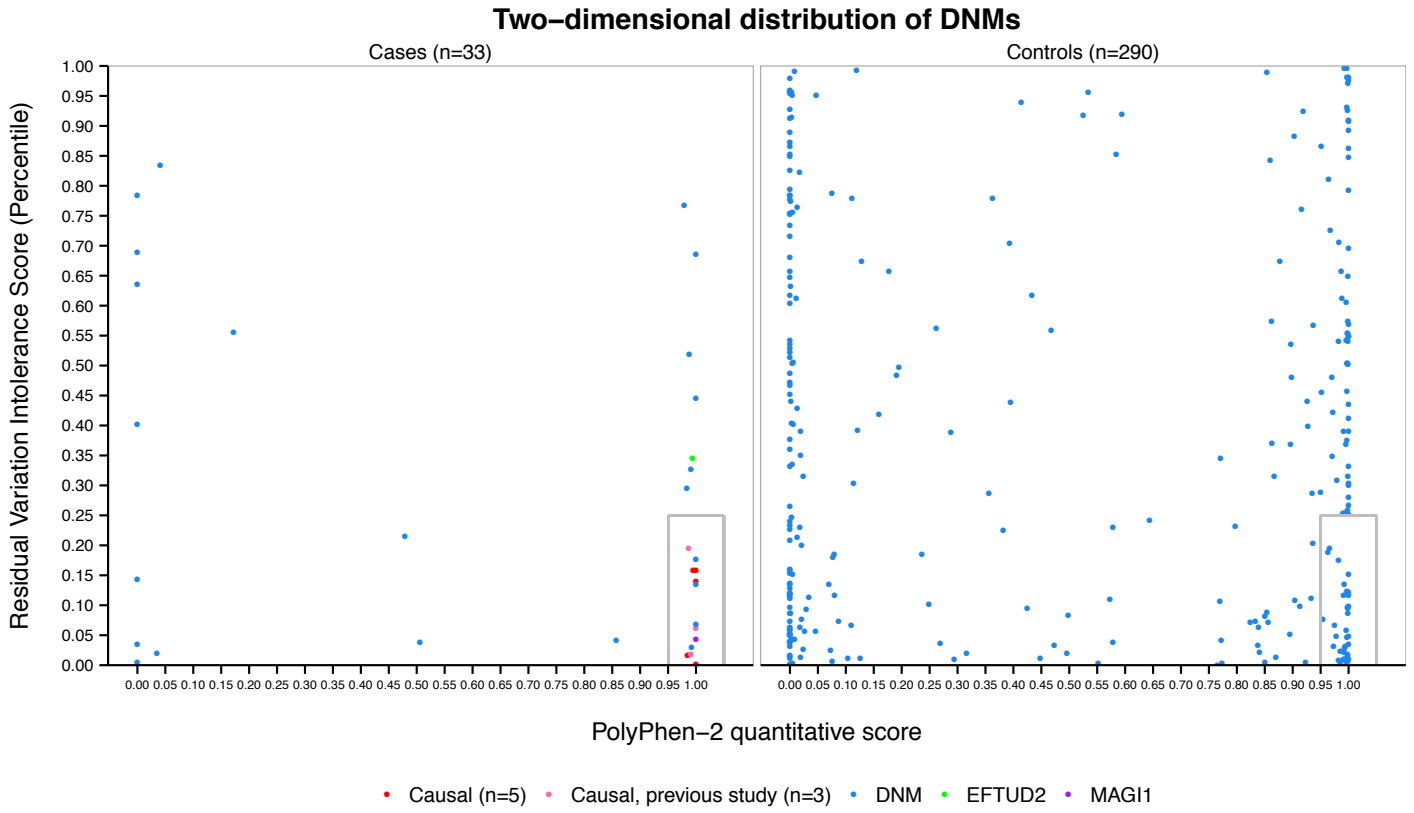


Figure 27. Distribution of the gene-level and variant-level scores for *de novo* mutations observed in patients with undiagnosed genetic conditions vs. control samples.

The grey box marks the boundaries of the "hot-zone". Blue points represent the most damaging *de novo* mutation identified in each trio. For ten trios the most damaging *de novo* mutation is color coded (not blue) to provide additional information. Red points

represent disease-causing DNMs (n=5) and pink points represent disease-causing DNMs identified in our previous study (n=3). Despite plotting the DNM predicted to have the most deleterious impact, two trios are actually explained by a different mutation *(not plotted).* The green point represents a DNM in *MAGI1*, however, the disease-causing variant in this trio is actually a non-hot-zone DNM in *SMAD4* [121]. The purple point represents a DNM in *ZFP90*, however, the disease-causing variant in this trio is actually in the intron-exon boundary of *EFTUD2* [121] and thus was outside of the CCDS coding regions considered in this analysis.

Table 15. *De novo* mutations with extreme gene-level score (RVIS ≤ 25%) and variant-level scores (PolyPhen-2 quantitative score ≥ 0.95).

[a]In addition to family 50, another hot zone *de novo* mutation (N266H) was seen in *SLC9A1* in a patient not included in this study that presented with spastic diplegia, autism, seizures, ID, behavioral problems and possible Worster-Drought syndrome. The phenotype was not similar enough to patient 50 to declare this as causal in either case. [b]N12 had two *de novo* mutations in the hot zone, of which, *SRCAP* was the more damaging of the two (methods), thus only *SRCAP* included in the 2D plot. Evidence for disease-associations based on previous studies: *KCNQ2*[144], *ASXL3*[137], *ATP8A2*[145], *GNAO1*[136], *KCNT1*[133,146], *HNRNPU*[42], *SRCAP*[135].

Trio	Variant location (GRCh37)	Gene	PolyPhen-2 (score)	RVIS (Percentile)	Disease-association based on previous studies
8	chr2:127826559	*BIN1*	probably damaging (1)	13.6	Unknown
10	chr20:62071037	*KCNQ2*	probably damaging (0.995)	15.9	Causal
33	chr18:31323963	*ASXL3*	frameshift variant (1)	15.9	Causal
41	chr13:26127983	*ATP8A2*	nonsense (1)	6.9	Maybe causal
46	chr16:56226491	*GNAO1*	probably damaging (1)	13.9	Causal
50[a]	chr1:27432470	*SLC9A1*	probably damaging (0.992)	2.9	Unknown
56	chr9:138669220	*KCNT1*	probably damaging (1)	1.6	Causal
N6	chr1:245020159	*HNRNPU*	Splice site variant (1)	17.6	Maybe causal
N12	chr16:30748691	*SRCAP*	nonsense (1)	0.2	Causal
N12[b]	chr11:47290219	*NR1H3*	probably damaging (1)	16.4	Unlikely

3.3.3 A PPI network approach for identifying candidate disease-causing mutations

For the 42 trios without a genetic diagnosis determined (Table 14), we took all genes carrying qualifying variants in each patient and used a PPI network approach to identify the direct binding partners for each of the proteins these genes encode. This identified between zero and 544 binding partner proteins per patient, with an average of 147 per patient. Inspection of the relevant genes and any associated phenotypes identified two good candidate variants in two patients with phenotypically overlapping features. The patient in family 28 carried a rare homozygous missense mutation in *MED29* (p.Leu139Pro), which is a direct binding partner of *MED17*. A homozygous missense mutation in *MED17* was previously reported as the cause of an almost identical phenotype [147]. *MED17* and *MED29* are both components of the Mediator complex. Therefore, the missense mutation in *MED29* is a good candidate in this patient.

The patient from family 42 suffers from severe intractable seizures with developmental delay and carries a qualifying compound heterozygous genotype in *ANK3*. The ANK3 protein directly binds to SCN2A [148] and SCN1B [149]. *SCN2A* is a known causal gene for epileptic encephalopathy [150], and mutations in *SCN1B* are the causes for generalized epilepsy with febrile seizures [151]. Both variants in *ANK3* were missense and predicted to be probably damaging by PolyPhen-2. *ANK3* mutations are implicated in several neurodevelopmental and psychiatric disorders[152]. The association between ANK3 and epilepsy is yet to be established, however since ANK3

directly binds two established epilepsy gene products, we considered this compound heterozygote to be good candidate in this patient.

3.4 Discussion

In this study of 62 patients with unresolved genetic conditions (56 newly sequenced, six previously reported but unresolved) we identified causal mutations in 32% of the patients based on observing qualifying variants in previously established and clinically relevant genes. Applying a gene- and variant-level prioritization framework to prioritize *de novo* mutations we found nine damaging *de novo* mutations in intolerant genes: five of which were considered causal mutations based on their location within previously established and clinically relevant genes. A protein-protein interaction analysis, seeded by genes carrying qualifying variants, identified candidate genetic diagnoses for two additional patients. By reanalyzing six previously unresolved cases, we were able to resolve two of them. For both of these, the causal variant was previously not called by the original variant calling program. In total, the expanded analyses identified 12 good candidate diagnoses beyond the 20 secure ones observed in already known genes.

While traditional clinical genetic screening failed to identify any of these disease-causing mutations, we obtained an overall molecular diagnostic rate of 32% (20 of the 62 studied patients). There are two major reasons that traditional clinical genetic screening was unsuccessful. First, many genetic disorders have been reported in only a few cases

or the causes of the genetic disorders have been identified only very recently. Second, for some phenotypes, the list of potential causal genes is large, thus it is not feasible to Sanger sequence all the candidate genes in a clinical lab or a clinical test may not be available. These analyses are consistent with the growing literature that suggests that exome sequencing is a viable alternative to traditional genetic testing, especially when the initial clinical evaluation and/or testing does not yield a diagnosis [153].

Our analyses suggest that exome sequence permits analyses that go beyond what is traditionally done in diagnostic settings (look for mutations in established disease genes). We saw that a PPI based method provided good candidates for two patients, which can now be kept on a "watch list" for implication in future research studies and/or followed up with molecular biology or animal models. Perhaps most encouraging is the clear indication that a combination of gene- and variant-level prioritization provides direct pointers to candidate causal mutations that is completely independent of a gene's previous association to disease. A clear example is the *GNAO1* causal diagnosis, which was reported in literature [136] after our analyses highlighted it within the hot-zone. This suggests that other hot-zone *de novo* mutations of unknown clinical significance may eventually be declared causal after gene-disorder associations become secure with studies in additional patients.

We observed that the presence of a deleterious *de novo* mutation in an intolerant gene was enriched 4-fold in patient exomes compared to control exomes (Figure 27). In

the newly sequenced probands whose phenotypes were not explained by inherited disease causal variants, 28 probands had at least one protein-coding *de novo* mutation and nine had hot-zone *de novo* mutations. While five of these *de novo* mutations were already determined to be disease-causing, the other four are novel candidates. In fact, two of these de novo mutations are in genes associated with similar phenotypes. However, additional evidence is needed to confirm the role of these variants.

There are many advantages to adopting exome sequencing in medical genetics clinics. First, it is fast. From the initiation of sequencing to obtaining finalized analyses often takes as little as a month, with the quickest turnaround time so far observed in our Center being 12 days. Indeed, a recent study reported that whole-genome sequencing could be done in 50 hours in a neonatal intensive care unit [154]. Rapidly attaining genetic diagnoses is important for appropriate treatment. For example, although the causal gene (*GNE*) of Sialuria has been known since 1999 [131], the patient from family 34 did not get a diagnosis for 11 years due to the extreme rarity of the condition, but was immediately recognized as having this condition once the *GNE de novo* mutation was observed. Immediately after we identified his causal mutation, he was seen by a world's leading expert on Sialuria to receive the appropriate management of his condition. Second, the cost is moderate. The cost of exome sequencing a trio is cheaper than the combined costs of multiple traditional genetic tests [153]. The cost of exome sequencing will continue to decrease with the evolution of sequencing technology and analysis

methods. Third, for patients without obvious causal mutations after exome sequencing analysis, data can be reanalyzed using newer methods that may lead to identification of causal mutations in the future. For example, in this study we revisited six unresolved trios [121]. We used Genome Analysis Toolkit to improve the sequencing reads alignment quality and to update the variant calling (previously SAMtools); and after reanalysis, two of the six patients were resolved by previously unidentified variants. Finally, qualifying variants in genes that are not currently reported in the literature as pathogenic can be periodically rechecked against newly published literature, to see if the gene has been newly implicated in disease.

While it is clear that exome sequencing cannot identify all causal mutations, the diagnostic yield is already remarkably high and will only continue to grow as the full phenotypic spectrums associated with mutation in individual genes are clarified and as analytic approaches continue to improve.

4. *De novo* mutations in epileptic encephalopathies[1]

4.1 Introduction

Epileptic encephalopathy describes a subset of epilepsy patients/syndromes presenting with severe cognitive and behavioral impairments that are thought to be a direct result of the epileptic activity itself[19]. Their clinical presentation lies on a spectrum of severity and these patients often progress, worsening over time and making early intervention especially critical for these patients. Epileptic encephalopathies include several severe childhood epilepsy disorders for which the cause is frequently unknown[19]. Given the recent discovery that *de novo* mutations contributed to disease risk for intellectual disability, autism spectrum disorder, and other severe neurodevelopmental syndromes [57-59,118], we investigated the role of *de novo* mutation in two "classical" types of epileptic encephalopathies: infantile spasms (IS) and Lennox-Gastaut syndrome (LGS). In some patients, infantile spasms evolve to Lennox-Gastaut syndrome.

While I was involved in many aspects of this project, my primary contributions included: sample organization and management, creation of a database for sample tracking with daily updates (accessible by all collaborators), IGV visual inspections, co-management of Sanger sequencing validation, inspection and organization of Sanger sequence traces, providing counts of confirmed *de novo* mutations in various categories

[1] This is part of a published work [42] involving the Epi4K and EPGP consortia.

(CCDS SNVs, SNVs in intolerant genes, and so on) to feed into all downstream analyses, identification of recessive variants, generation of the high confidence list of genes previously associated with epileptic encephalopathy, and writing of the manuscript[42].

4.2 Materials and methods

4.2.1 Subjects

Infantile spasms and Lennox–Gastaut syndrome patients evaluated in this study were collected through the Epilepsy Phenome/Genome Project (EPGP)[155]. The appropriate Institutional Review Boards approved the use of samples in this research.

The EPGP team enforced strict diagnostic and exclusion criteria for inclusion of these samples[42]. These included characteristics for patient electroencephalogram (EEG) readings, absence of developmental delay or ASD prior to seizure onset, absence of any obvious structural (including MRI screenings) or metabolic causes [42]. This resulted in a cohort of 264 patients, comprised of epileptic encephalopathy patients (n = 149) and Lennox–Gastaut syndrome (n = 115), and their unaffected biological parents.

This study also used two control cohorts to assess variant frequencies: (i) 436 unrelated controls exome-sequenced in the Center for Human Genome Variation as part of other genetic studies and (ii) ~6,500 subjects was available from the National Heart, Lung, and Blood Institute (NHLBI) Grand Opportunity (GO) Exome Sequencing Project (ESP) for the identification of genes contributing to heart, lung, and blood disorders (Exome Variant Server [EVS], NHLBI ESP, Seattle, WA).

Finally, we make use of published *de novo* mutation data from trios using unaffected siblings of ASD patients as controls[62,63,125].

4.2.2 Exome sequencing and *de novo* mutation identification

DNA from the parents and probands were used to create sequencing libraries and the Illumina TruSeq Exome Enrichment kit was used to selectively amplify the targeted exonic and flanking intronic regions of the genome. All individuals from a trio (the patient/proband and their unaffected biological parents) were always sequenced simultaneously to avoid batch effects; with two complete trios sequenced in parallel (six barcoded samples) across two lanes of an Illumina HiSeq 2000 sequencer.

The resulting short-sequence reads were aligned to the reference genome (NCBI Build 37) using the Burrows-Wheeler Alignment (BWA, version 0.5.10) tool[83] and PCR duplicates were removed using the Picard software (http://picard.sourceforge.net).

The GATK Unified Genotyper was used to jointly call putative *de novo* mutations in the proband, based on all family members in a trio. Subsequently, putative de novo mutations were extracted from the variant call format files (VCFs) that met the following criteria: (1) the read depth in both parents was ≥10; (2) the depth of coverage in the child was at least one-tenth of the sum of the coverage in both parents; (3) < 5% of the reads in either parent supported the alternate allele/*de novo* allele; (4) ≥25% of the reads in the

child carried the alternate allele; (5) the PL scores[2] for the offspring genotypes AA, AB, and BB, were >20, 0, and >0, respectively; (6) the PL scores for both parental genotypes were 0, >20, and >20; (7) at least three variant alleles were observed in the proband; and (8) the *de novo* variant had to be located in a CCDS exon targeted by the exome enrichment kit.

SnpEff (version 3.0a) was used to annotate the variants according to Ensembl (version 69) and consensus coding sequencing (CCDS release 9, GRCh37.p5) and all subsequent analyses were limited to exonic or splice site (2 bp flanking an exon) mutations.

4.2.3 *De novo* mutation validation

All putative *de novo* mutations that were absent from both control cohorts were also visually inspected using Integrative Genomics Viewer (IGV) to eliminate any variants resulting from poor alignment. All putative *de novo* mutations were then Sanger sequenced in the relevant proband and both parents. For comparison, we also individually called *de novo* variants from probands and parents (GATK) for a subset of trios. Using the individual variant calling approach for putative *de novo* mutation identification, we subsequently confirmed an additional 46 *de novo* mutations. These were included in all the downstream *de novo* mutation analyses.

---

[2] The normalized, phred-scaled likelihood (PL) scores are assigned such that the most likely genotype is given a score of 0, and the score for the other two genotypes represent the likelihood that they are not the true genotypes. For the genotypes AA, AB, and BB: A is the reference allele and B is the alternate allele.

For the 264 exome sequenced epileptic encephalopathy trios, the sequenced DNA-source was derived from primary cells in 224 trios or from lymphoblastoid cell lines (LCLs) in one or more family members in 40 trios. In all cases, primary DNA from the proband was used for the Sanger confirmation thus eliminating any mutations appearing as a result of the transformation process for the 40 trios sequenced from LCLs.

4.2.4 Defining the opportunity space for detecting *de novo* mutations

For each trio, we defined exonic bases that were sufficiently covered (≥10x) in all three family members and thus had the opportunity for identification of a *de novo* mutation. Specifically, we restricted to bases in the consensus coding sequence (CCDS release 9, GRCh37.p5) or within the two base pairs at each end of exons to allow for splice acceptor and donor variants. Additionally, we required a raw phred-scaled genotype confidence score of ≥20 (regardless of the presence or absence of a variant) obtained after multisample variant calling (GATK). Using these three criteria, the average CCDS-defined *de novo* mutation opportunity space across 264 trios was found to be 28.84 ± 0.92 Mb (range of 25.46–30.25 Mb).

Similarly, for each trio, we assessed the *de novo* calling opportunity space for every gene with a CCDS transcript. For genes with any non-overlapping CCDS transcripts, we merged the corresponding regions into a consensus summary of all CCDS-defined bases for that gene. Under the CCDS opportunity space criteria, over 85% of the CCDS-defined exonic regions were sufficiently covered (≥10x) across all three

family members in over 90% of the trios and all 264 trios covered at least 79% of the CCDS-defined regions.

4.2.5 Calculation of gene-specific mutation rates

Point mutation rates were scaled to per base pair, per generation, based on the human genome sequences matrix[156] (provided by S. Sunyaev and P. Polak), and the known human average germline *de novo* mutation rate (1.2×10^{-8} per base pair per generation) [157]. After restricting to the CCDS opportunity space, the mutation rate (M) of each gene was calculated by adding up the point mutation rates in the probands, and then dividing by the total trio number ($S = 264$). The P value for each gene was calculated as [1 – Poisson cumulative distribution function (x - 1, λ)], where x is the observed *de novo* mutation number for the gene, and λ is calculated as $2SM$ for genes on the autosomes or $(2f + m)M$ for genes on the X chromosome (where, f and m are the number of sequenced female and male probands, respectively). Genes on Y chromosome were not part of these analyses. Two distinct probands carry identical *de novo* mutations in *ALG13* and two different probands carry identical *de novo* mutations in *SCN2A*. We calculated the probability of this special case as [1 – Poisson cumulative distribution function (1, $(2f + m)r$)], where r reflects the point mutation rate for that specific *de novo* mutation position. Further investigations suggested that the occurrence of these identical *de novo* mutations in distinct probands was unlikely to have been caused by sequencing or mapping errors (data not shown)[42].

4.2.6 Application of the gene-specific mutation rate calculation to genes previously associated with epileptic encephalopathy

We compiled a comprehensive list of putative epileptic encephalopathy genes (n=52) by curating of the OMIM database [January 2014] as well as several recent publications. We considered five categories of genes: (I) OMIM "epileptic encephalopathy (EIEE)", n=18; (II) OMIM "Dravet Syndrome", n=1; (III) OMIM Neurological Clinical Synopsis containing "epileptic encephalopathy" in disorders with a "known molecular basis", n=8; (IV) Genes implicated in OMIM by the appearance of "epileptic encephalopathy" in the gene summaries, n=19; (V) those implicated in epileptic encephalopathy in the recent literature, n=5 (*ALG13*[42], *GABRA1*[158], *GABRB3*[42], *GRIN2B*[159], and *SLC35A2*[61]).

For the first two categories, if the mode of inheritance was recessive or if the first reported allelic variant in OMIM was published before 2010, we considered these definitive epileptic encephalopathy genes (n=15; *ARHGEF9, ARX, CDKL5, KCNQ2, PCDH19, PLCB1, PNKP, PNPO, SCN1A, SCN2A, SLC25A22, ST3GAL3, STXBP1, SZT2, TBC1D24*). For all remaining genes on this comprehensive list of putative epileptic encephalopathy genes, a panel of clinicians reviewed the specific associated phenotypic details to determine the relevance to epileptic encephalopathy specifically. This highlighted 12 genes of interest, including: eight genes from category I, two genes from category III, one from category IV, and five genes from category V.

We then collected details from the primary literature for each of these 12 genes, including both positive and negative studies in order to assess the full opportunity to observe a mutation. We assumed full coverage of all targeted genes in a given study. Singleton case studies and studies with non-EE phenotypes were excluded. Only single nucleotide variants and *de novo* mutations were considered. Only indels have been reported in the *SPTAN1* gene and thus no qualifying mutations were available for analysis.

Additionally, only confirmed *de novo* mutations were considered and only complete trios were counted toward the full cohort size. If a reference reported all detected *de novo* and "possibly *de novo*" (lack of parental DNA) then the total cohort size was counted regardless of the number of complete trios, and only confirmed de novos counted towards the observed *de novo* mutation counts. However, we included one variant where a "possibly *de novo* mutation" (lack of parental DNA) because the exact same mutation was seen as a confirmed *de novo* mutation in another case (*GABRA1)*[158] and one variant with mosaicism in the father that was germ line in the child (*SCN8A*) [41].

For one variant (12_52200572_A/G) in *SCN8A*, this exome-sequenced trio was originally withheld from the publication[42] because the same exact variant was identified in another study[160] and we were concerned that it was the same patient; for the purposes of this exercise we added this trio back into the evaluation of all 12 genes.

4.3 Results

4.3.1 Distribution of *de novo* mutations

Exome sequencing of 264 trios led to the identification of 439 putative *de novo* mutations. Sanger sequencing confirmed 329 *de novo* mutations, and the remainder were either false positives, a result of B-cell immortalization, or in regions where the Sanger assays did not work.

On average each trio harbored 1.25 confirmed *de novo* mutations, with 181 probands harboring at least one (Figure 28). Considering only SNVs, on average, each trio harbored 1.17 *de novo* mutations. Seventy-two per cent of the confirmed *de novo* mutations were missense and 7.5% were putative loss-of-function (splice donor, splice acceptor, or stop-gain mutations). Compared to rates of these classes of mutations previously reported in controls (69.4% missense and 4.2% putative loss-of-function mutations)[62,63,125], we observed a significant excess of loss-of-function *de novo* mutations in infantile spasm and Lennox–Gastaut syndrome patients (exact binomial P = 0.01). A similar excess of loss-of-function *de novo* mutations has been reported in autism spectrum disorders[58,62,63,125].

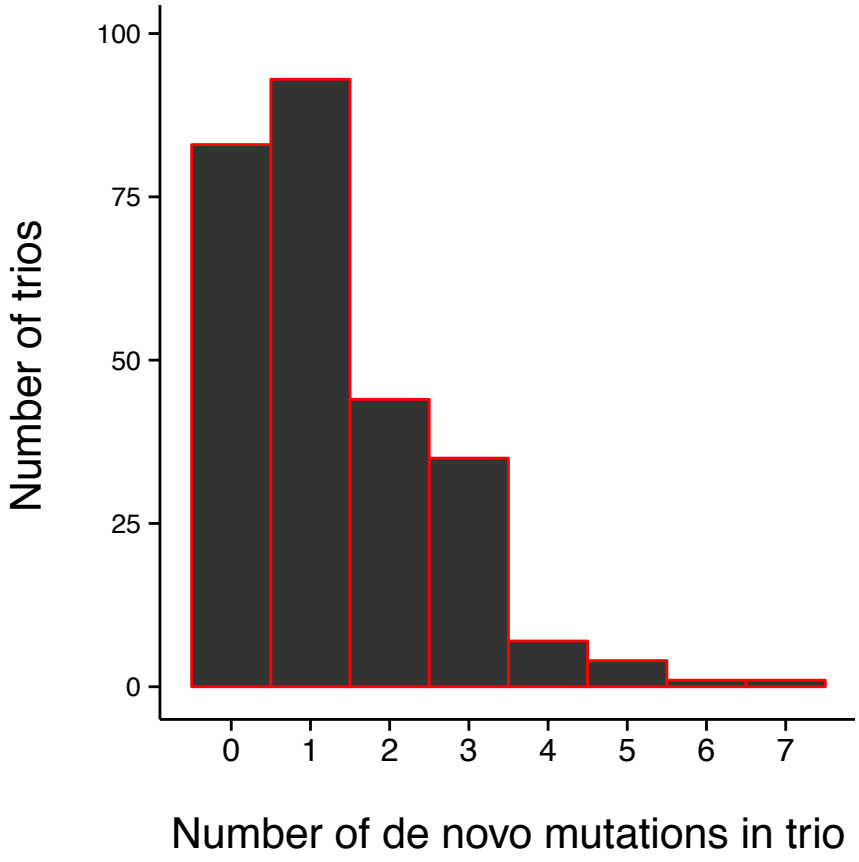


Figure 28. Distribution of *de novo* mutations detected in 264 IS/LGS probands. Histogram is based on all Sanger sequencing confirmed *de novo* mutations (n=329) from the 264 IS/LGS trios.

4.3.2 Recurrently mutated genes in these epileptic encephalopathy patients

Across the 264 trios, we found nine genes with *de novo* SNV mutations in two or more probands (Figure 29). Five of these *de novo* mutations are found in previously established epileptic encephalopathy genes: *SCN1A* (MIM#607208), *STXBP1* (MIM#612164), *SCN8A* (MIM#614558), *SCN2A* (MIM #613721), and *CDKL5* (MIM

#300672). The remaining four genes have no known association with epileptic encephalopathies: *GABRB3*, *ALG13*, *DNM1*, and *HDAC4 (*Figure 29).

To assess whether the observations in these genes implicate them as novel risk factors for epileptic encephalopathies, we determined the probability of seeing multiple mutations in the same gene given the sequence-specific mutation rate, size of the gene, and the number and gender of patients evaluated in this study. This calculation revealed that the number of observed *de novo* mutations in *HDAC4* and *DNM1* are not significantly greater than the null expectation. However, observing four unique *de novo* mutations in *GABRB3* and two identical *de novo* mutations in *ALG13* were found to be highly improbable; providing clear statistical evidence of an association with epileptic encephalopathy (Table 16 and Figure 30). For comparison, we performed the same calculation for genes with multiple *de novo* mutations observed in 610 control trios[62,63,125], and found no genes with a significant excess of *de novo* mutations (data not shown).

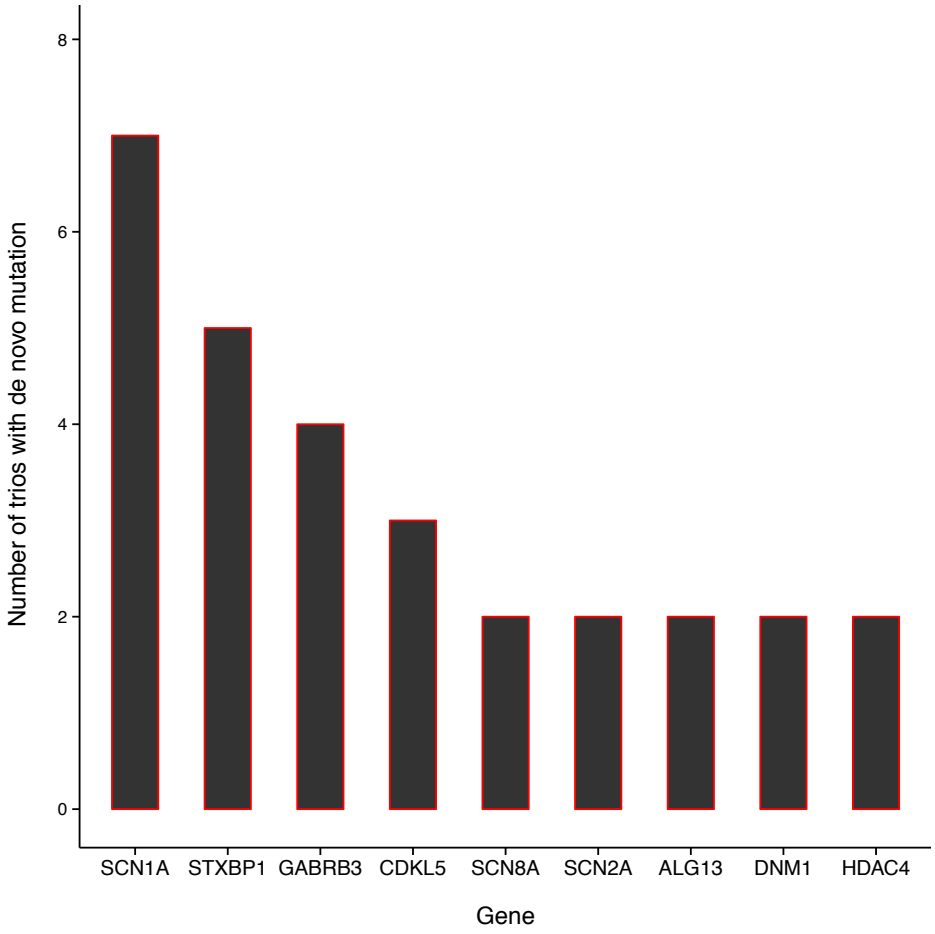


Figure 29. Number of trios with a *de novo* mutation in recurrently mutated genes.

Table 16. Probability of observing the reported number of *de novo* mutations by chance in genes recurrently mutated in this cohort

[a]Adjusted α is equivalent to 0.05/18,091 = 2.76 X 10^{-6} (*), 0.01/18,091 = 5.53 X 10^{-7} (**) and 0.001/18,091 = 5.53 X 10^{-8} (***). [b]Count for *SCN1A* excludes three additional patients with an indel or splice site mutation as these are not accounted for in the mutability calculation. [c]Two *de novo* mutations occur at the same position. The probability of these special cases obtain P = 7.77 X 10^{-12} and P =1.14 X 10^{-9} for *ALG13* and *SCN2A*, respectively.

Gene	Chr	Average effectively captured length (bp)	Gene-specific mutation rate	Observed *de novo* mutation number	*P* value[a]	
SCN1A	2	6,064	1.61 X 10^{-4}	5[b]	1.12 X 10^{-9}	***
STXBP1	9	1,918	6.44 X 10^{-5}	5	1.16 X 10^{-11}	***
GABRB3	15	1,207	3.78 X 10^{-5}	4	4.11 X 10^{-10}	***
CDKL5	X	2,798	5.44 X 10^{-5}	3	4.90 X 10^{-7}	**
ALG13[c]	X	475	1.03 X 10^{-5}	2	7.77 X 10^{-12}	***
DNM1	9	2,323	9.10 X 10^{-5}	2	2.84 X 10^{-4}	
HDAC4	2	2,650	1.16 X 10^{-4}	2	4.57 X 10^{-4}	
SCN2A[c]	2	5,831	1.52 X 10^{-4}	2	1.14 X 10^{-9}	***
SCN8A	12	5,814	1.64 X 10^{-4}	2	9.14 X 10^{-4}	

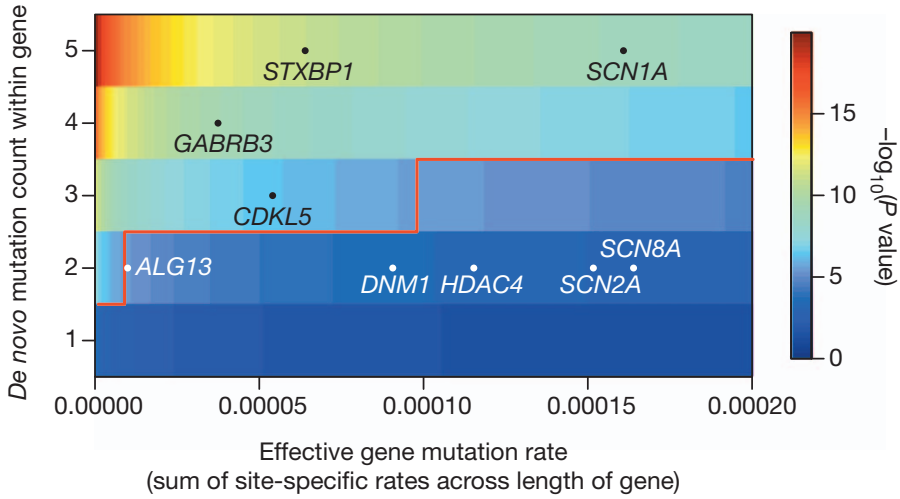


Figure 30. Heat map illustrating the probability of observing the specified number of *de novo* mutations in genes with the specified estimated mutation rate. The number of *de novo* mutations required to achieve significance for a gene with the specified gene-specific mutation rate is indicated by the solid red line. The superimposed dots reflect positions of genes recurrently mutated in our study. *GABRB3*, *SCN1A*, *CDKL5* and *STXBP1* have significantly more *de novo* mutations than expected. The positions indicated for *ALG13* and *SCN2A* reflect only the fact that there are two mutations observed, ignoring that there are two mutations affecting the same site.

4.3.3 Distribution of *de novo* mutations in intolerant genes

In a trio-based sequencing study of ASD, a simulation-based framework was developed for testing whether the distribution of *de novo* mutations in affected individuals differed from that of the general population[58]. We extended this by developing a likelihood model that describes the distribution of *de novo* mutations among affected individuals in terms of the distribution in the general population and a set of parameters describing the genetic architecture of the disease[42].

To describe the genetic architecture of the disease, in this case epileptic encephalopathy, we included several parameters[42]. Briefly, these parameters included: the proportion of the exome sequence that can carry disease-influencing mutations (η) and the relative risk (γ) of the mutations. Consistent with what was reported in autism spectrum disorder[58], we found no significant deviation in the overall distribution of mutations from expected ($\gamma = 1$ and/or $\eta = 0$).

Given that some genes are more tolerant to protein-disrupting mutations in that they do not result in overtly adverse phenotypic consequences[124], we assessed the RVIS values [124] for genes with a known association with epileptic encephalopathy. We found that these known epileptic encephalopathy genes (n=14, Table 17) rank among the most intolerant genes in the genome, with 12 of the 14 genes falling in the lowest 25^{th} percentile (the most intolerant and lowest RVIS values). We therefore evaluated the distribution of *de novo* mutations observed in our epileptic encephalopathy trios, within the 4,264 genes that are within the 25th percentile for intolerance and found a significant shift from the null distribution ($P = 2.9 \times 10^{-3}$). The maximum likelihood estimates of η (percentage of intolerant genes involved in epileptic encephalopathies) was 0.021 and γ (relative risk) was 81, indicating that there are 90 genes among the intolerant genes that can confer risk of epileptic encephalopathies and that each mutation carries substantial risk.

Table 17. The opportunity space to call a *de novo* variant in the known early epileptic encephalopathy MIM genes and their intolerances scores (RVIS). The identification of genes with a known association with epileptic encephalopathy is based solely on OMIM® [accessed December 2012]. *ARX* was excluded because it had < 70% of the gene covered.

Phenotype MIM number	OMIM Phenotype	Gene	Average % of CCDS transcript with *de novo* mutation call opportunity	RVIS score and (percentile)
300088	Epileptic encephalopathy, early infantile, 9	*PCDH19*	91.80%	-0.89 (10.4%tile)
300607	Epileptic encephalopathy, early infantile, 8	*ARHGEF9*	87.64%	-0.12 (44.9%tile)
300672	Epileptic encephalopathy, early infantile, 2	*CDKL5*	88.20%	-0.67 (15.9%tile)
308350	Epileptic encephalopathy, early infantile, 1	*ARX*	20.57%	*Unassessed*
606369	Epileptic encephalopathy, Lennox-Gastaut type	*MAPK10*	99.76%	-0.45 (24.0%tile)
607208	Dravet syndrome	*SCN1A*	98.86%	-1.43 (4.0%tile)
609304	Epileptic encephalopathy, early infantile, 3	*SLC25A22*	81.59%	0.13 (63.2%tile)
612164	Epileptic encephalopathy, early infantile, 4	*STXBP1*	97.09%	-0.69 (15.0%tile)
613402	Epileptic encephalopathy, early infantile, 10	*PNKP*	86.67%	-0.53 (20.9%tile)
613477	Epileptic encephalopathy, early infantile, 5	*SPTAN1*	98.56%	-3.53 (0.3%tile)
613720	Epileptic encephalopathy, early infantile, 7	*KCNQ2*	56.04%	-0.67 (15.9%tile)
613721	Epileptic encephalopathy, early infantile, 11	*SCN2A*	93.78%	-1.99 (1.8%tile)
613722	Epileptic encephalopathy, early infantile, 12	*PLCB1*	97.05%	-0.84 (11.5%tile)
614558	Epileptic encephalopathy, early infantile, 13	*SCN8A*	96.15%	-1.75 (2.3%tile)
614959	Epileptic encephalopathy, early infantile, 14	*KCNT1*	68.59%	-2.07 (1.6%tile)

4.3.4 Other highly penetrant genotypes in epileptic encephalopathy patients

In addition to *de novo* variants, we also screened for highly penetrant genotypes by identifying variants that create newly homozygous, compound heterozygous, or hemizygous genotypes in the probands that are not seen in parents or controls.

Focusing only on rare (control MAF <0.0015) and likely functional (missense, nonsense, or splice site) CCDS variants, there were 130 newly homozygous variants and 414 newly hemizygous variants in 370 unique genes, impacting158 trios. A newly hemizygous variant was found in *THOC2,* which harbors a confirmed *de novo* mutation

in a different trio. Moreover, two trios harbor newly homozygous variants and 24 trios harbor newly hemizygous variants in genes known to cause monogenic disorders that include seizures[43], including six trios with a newly hemizygous variant in *ATRX.*

Additionally, we considered genes where a rare and likely protein-disrupting variant was inherited from each parent (compound heterozygotes) and in which co-occurrence of alleles was completely absent in the CHGV control cohort. This included 561 compound heterozygous variants in 351 unique genes, five of which were found in genes known to cause monogenic disorders that include seizures[43].

No inherited variants showed significant evidence of association. Similar classes of mutations are not reported in trio control cohorts[62,63,125] making comparisons to expectation difficult. Additional studies evaluating a larger number of epileptic encephalopathy patients will be required to establish the role of inherited variants in the disease risk associated with infantile spasms and Lennox–Gastaut syndrome.

4.3.5 Cross-study validation of genes previously associated with epileptic encephalopathy

To generate a high confidence list of solved epileptic encephalopathy trios, we first curated the OMIM database and several recent publications (n=5; *ALG13*[42], *GABRA1*[158], *GABRB3*[42], *GRIN2B*[159], and *SLC35A2*[61]) to obtain a comprehensive list of putative epileptic encephalopathy genes. A subset of these genes (n=15), for example *SCN1A*, was considered definitive based on the volume and substantial evidence from previous studies (methods). Next, a panel of clinicians reviewed the

phenotypes of patients reported to carry a mutation in one of the remaining putative epileptic encephalopathy gene. If the phenotypic details were deemed consistent with epileptic encephalopathy, then these genes were statistically evaluated for the evidence of their association with epileptic encephalopathy. In total, we tested 12 genes.

To test these genes as was done in Table 16, we first curated the primary literature supporting the association between each of these 12 genes and epileptic encephalopathy, including both positive and negative studies in order to assess the full opportunity to observe a mutation. We assumed full coverage of all targeted genes in a given study. Singleton case studies and studies with non-epileptic encephalopathy phenotypes were excluded. Only validated SNV *de novo* mutations were considered (methods).

Three studies included exome-sequenced trios and thus these three studies counted towards the total cohort size for all 12 genes [42,146,158]. Similarly, we included a yet unpublished study of an additional 92 exome-sequenced epileptic encephalopathy trios from the Epi4K consortium. Finally, the total cohort sizes for individual genes varied if a resequencing study targeted one of these 12 genes.

Of the 12 tested genes, nine were significantly associated with epileptic encephalopathy. Together with the original 15 definitive epileptic encephalopathy

genes, we confirm 24 high confidence epileptic encephalopathy genes. A full list of contributing variants and literature references for each gene are provided in Table 18[3].

We then used this list to identify the proportion of epileptic encephalopathy trios that can be explained by a *de novo* mutation in one of these 24 genes. We find that in our cohorts, including the originally sequenced trios (n=264)[42] and our yet unpublished study of an additional 92 trios, *de novo* mutations in these genes explain ~12% of epileptic encephalopathy genes.

---

[3] Additional information for Table 18.
Five cohorts contribute to all genes: 92 new unpublished Epi4K trios*, the *SCN8A* trio previously excluded from[42], and three exome sequencing studies[42,146,158]. Studies were also included for specific genes: *ALG13*[42], *CHD2*[41,42,161], *GABRA1*[42,158], *GABRB3*[42], *GNAO1**[42,136], *GRIN2A*[61,162,163], *GRIN2B*[42,159], *KCNT1*[42,146] , *SCN8A*[41,42], *SLC35A2*[42] [164], and *SYNGAP1*[41],[165], [166].
[†]Adjusted α is equivalent to (0.05 / 18,091) = 2.76 x10^{-6}.
[#]Two *P* values were calculated, considering all individuals are females or males, respectively.
[!]These genes have multiple identical mutations occurring in different individuals. *P*-values for these special cases are in the table, and *P* values ignoring the fact that these are identical mutations are as follows: *ALG13* (5.97 x10^{-4}; 1.51 x10^{-5}), *GABRA1* (3.16 x10^{-9}), and *KCNT1* (3.47 x10^{-10}).

Table 18. Genes with *de novo* mutations reported in the literature and the probabilities of getting greater than or equal observed *de novo* mutation numbers by chance.

Gene	Chr	Length (bp)	Gene-specific mutation rate	*De novo* mutation carrier count	*De novo* mutation details	Mutation negative non-carrier count	p-value †	
ALG13 [#!]	X	3,641	4.73×10^{-5}	2	X_110928268_A/G, X_110928268_A/G	368	1.75×10^{-11}; 4.38×10^{-12}	***
CHD2	15	5,643	7.68×10^{-5}	8	15_93496726_T/C, 15_93515610_T/C, 15_93470540_C/T, 15_93492307_G/A, 15_93472268_C/T, 15_93563306_G/A, 15_93499689_A/C, 15_93492200_C/T	1,021	8.43×10^{-12}	***
GABRA1 [!]	5	1,407	1.88×10^{-5}	5	5_161317951_G/A, 5_161300202_G/A, 5_161300202_G/A, 5_161322732_A/C, 5_161322690_C/T	435	1.01×10^{-11}	***
GABRB3	15	1,573	2.66×10^{-5}	4	15_26866594_T/C, 15_26866564_C/T, 15_26828484_T/C, 15_26806254_T/C	366	6.12×10^{-9}	***
GNAO1	16	1,447	2.77×10^{-5}	5	16_56385396_T/C, 16_56368697_A/G, 16_56370656_G/A, 16_56374858_T/A, 16_56385380_A/C*	651	5.12×10^{-10}	***
GRIN2A	16	4,443	7.47×10^{-5}	4	16_9923442_C/A, 16_9916208_A/G, 16_9923333_A/G, 16_9934513_C/T	518	1.45×10^{-6}	**
GRIN2B	12	4,503	7.60×10^{-5}	3	12_13768545_C/A, 12_13761694_A/C, 12_13761703_T/A	1,005	5.34×10^{-4}	
KCNT1 [!]	9	3,832	8.73×10^{-5}	7	9_138667192_C/G, 9_138660694_G/A, 9_138671275_G/A, 9_138657552_G/A, 9_138657552_G/A, 9_138657552_G/A, 9_138671246_C/T	872	6.68×10^{-13}	***
SCN8A	12	6,047	8.86×10^{-5}	4	12_52180374_C/G, 12_52082568_G/A, 12_52159534_T/A, 12_52200572_A/G	866	2.08×10^{-5}	
SLC35A2 [#]	X	1,367	2.54×10^{-5}	3	X_48762548_G/A, X_48762503_G/T, X_48762684_G/A*	700	7.37×10^{-6}; 9.33×10^{-7}	***
SPTAN1	9	7,658	1.23×10^{-4}	0	n/a	431	1	
SYNGAP1	6	4,108	7.34×10^{-5}	5	6_33400486_A/T, 6_33408564_C/T, 6_33409537_G/A, 6_33405482_G/A, 6_33400501_C/T	669	7.27×10^{-8}	***

4.4 Conclusions

We identified *GABRB3* and *ALG13* as two novel genes associated with epileptic encephalopathy. Additionally, we describe a genetic architecture that strongly suggests that we have identified additional causal mutations in the subset of genes that are intolerant (25th percentile) to functional variation. Given that we see additional genes with recurrent mutations, it is likely that even modest increases in sample sizes will confirm many new genes now seen in only a single trio. However, the fact that we now know of 24 epileptic encephalopathy genes which explain only ~12% of patients emphasizes that the epileptic encephalopathies are genetically highly heterogeneous. Large cohorts of well-characterized patients will be needed to identify additional genes and add evidence to genes now implicated in only a single trio.

Two additional areas of research in this project [42] were the characterization of the observed genotype-phenotype relationships and an analysis of protein–protein interaction networks. The clinical characterizations revealed that it will be difficult to predict the responsible gene, even among known genes, based upon clinical presentation. For example, some genes harbor *de novo* mutations in both IS and LGS patients. To some degree, this is not surprising as it is known that some individuals with IS have a clinical course that evolves into LGS. However, we also see overlap with intellectual disability or autism spectrum disorders with patients also having *de novo* mutations in genes implicated in our cohort. This heterogeneity suggests that targeted

diagnostic sequencing will frequently be more laborious (with multiple negative genes being sequenced sequentially) than focusing on the genome as a whole. The protein–protein interaction analysis found that the genes carrying *de novo* mutations in these epileptic encephalopathy patients were highly interconnected, suggesting that many of these mutations converge on specific biological pathways, which would likely be beneficial for both drug development and personalized treatment.

5. Identification of epilepsy susceptibility variants in multiplex families

5.1 Introduction

To identify genetic factors influencing epilepsy susceptibility, whole-genome or exome sequencing was used to interrogate genetic variation in epilepsy cases from ~40 multiplex families. We focused on multiplex epilepsy families under the hypothesis that these cases would be enriched for genetic control. Our main hypothesis was that, in at least some proportion of these families, a single variant would explain all instances of epilepsy. To test this hypothesis, functional variants that were present in the epilepsy case genomes, but at very low frequencies in control genomes, were considered as candidate variants. If more than one individual per family was sequenced, then candidate variants were shared by all affected family members. Candidate variants were further assessed by cosegregation testing, variant association testing in a case-control cohort, and gene-based resequencing in a cohort of additional multiplex epilepsy families.

Two familial epilepsy discovery cohorts were used for this research. Discovery Cohort A is comprised primarily of families with no recognized syndromic presentation, while Discovery Cohort B (and Replication Cohort D) contains many families with familial temporal lobe epilepsy or the established epilepsy syndrome, Generalized Epilepsy with Febrile Seizures Plus (GEFS+). Pathogenic epilepsy CNVs are very rare in controls, suggesting that they have relatively high penetrance and these CNVs do not

exhibit phenotypic specificity in that they are found in multiple epilepsy syndromes and seizure types. Therefore, we hypothesized that other rare variants (SNVs and indels) may have similar properties in terms of their impact on disease risk. Such causal variants would be consistent with the range of phenotypes observed across and within the families of Discovery Cohort A. In contrast, the known causal mutations for GEFS+ families show Mendelian patterns of inheritance. However, given the large number of unexplained GEFS+ families, we highly suspect that additional novel genetic risk variants exist. Furthermore, the number of genes responsible for similar syndromes is not known. For these reasons, we were inclusive in the types of multiplex families we examined, looking at both syndromic and nonsyndromic presentations.

5.2 Materials and methods

5.2.1 Subjects

The appropriate Institutional Review Boards approved the use of each cohort in this research.

5.2.1.1 Discovery cohort A

The first discovery cohort consists of 88 epilepsy families collected by Dr. Ruth Ottman, at Columbia University. These families all have at least two affected individuals and on average, contain 3.8 (range 2-8) affected individuals per family with non-acquired epilepsy, and an average of 3.9 (range 2-9) individuals with non-acquired epilepsy or isolated unprovoked seizures. These families are not enriched for any

specific epilepsy syndromes, although a few of them have clinical characteristics consistent with a specific syndrome (e.g, GEFS+, Familial Adult Mycoclonic Epilepsy) or similar to a previously defined syndrome (e.g., Epilepsy, Familial Temporal lobe, 1 (MIM#600512) or GEFS+). More than half of the families consist of "mixed phenotypes" in that they contain some individuals with generalized epilepsy and others with focal epilepsy. In these families linkage analysis was performed using short tandem repeat polymorphisms spaced at an average of 9cM throughout the genome. Linkage from the four families presenting with GEFS+ was used to rule out the likelihood (no linkage peaks over known loci) that known genes explained these families. When assuming a model of 75% penetrance, no linkage peaks had LOD scores >3.0. Pedigrees for the 29 families where at least one sample was NGS, are provided in Appendix B.

5.2.1.2 Discovery cohort B

The second discovery cohort consists of 10 multiplex epilepsy families collected by Drs. Samuel Berkovic and Ingrid Scheffer at The University of Melbourne (Victoria, Australia). These families have an average of 9.7 affected individuals (excluding symptomatic epilepsies, individuals suffering from only febrile seizures, and diagnoses of suffering from a "possible seizure"). The families can broadly be described as follows: 2 focal epilepsy, 1 temporal lobe epilepsy (TLE), 4 generalized epilepsy with febrile seizures plus (GEFS+), 2 GEFS+ and idiopathic generalized epilepsy (IGE), and 1 IGE with additional complex phenotypes (i.e., occipital seizures).

Depending on the phenotypic diagnosis, these families were screened for phenotypically relevant genetic aberrations (Table 19). This screening did not identify a pathogenic genetic cause of epilepsy for all affected individuals in any of these families. Pedigrees for the 10 families where at least one sample was NGS, are provided in Appendix B.

Table 19. Genetic screening in the Discovery Cohort B families.

Family	Family ID	Previous screens (all negative for mutations)	Familial diagnosis
A	80389	LGI1, 15q13.3	Familial focal epilepsy
B	82051	LGI1, 15q13.3	Familial focal epilepsy
C	82654	LGI1, 15q13.3, SCN1A, SCN1B, GABRG2, SLC2A1	Familial temporal lobe epilepsy (TLE)
D	80553	SCN1A, SCN1B, 15q13.3, CACNA1A	GEFS+
E	82320	SCN1A, SCN1B, 15q13.3, CACNA1A, SLC2A1	GEFS+/IGE overlap
F	80780	15q13.3, CACNA1A, SLC2A1	GEFS+/IGE overlap
G	82439	15q13.3, CACNA1A, CACNA1H	GEFS+
H	83443	15q13.3, SLC2A1	IGE/complex phenotypes
I	83909	SCN1A, SCN1B, 15q13.3	GEFS+
J	80707	SCN1A, SCN1B, GABRG2, 15q13.3, CACNA1H, SLC2A1	GEFS+

5.2.1.3 Replication cohort C

This epilepsy cohort was collected as a collaborative effort between our lab (Center for Human Genome Variation (CHGV) at Duke University) and sites in Ireland, England, and Belgium. These sites came together in 2005 to form the Epigen consortium. As a member of the Epigen consortium, our lab has access to over 4,000 epilepsy samples; to our knowledge this is one of the largest collections of epilepsy samples in the world. Importantly, we have access to the original clinical records for all of these

patients. The Epigen cohort is not a family-based cohort, although there are a small number of families in this cohort.

The Epigen cohort can be used for follow up assessment of variants identified in the familial cohort. Therefore, knowing the phenotypic composition of this cohort is also useful because we hypothesize that some susceptibility variants may be enriched in specific epilepsy phenotypes. The Epigen patient cohort is approximately 3% absence epilepsy, 8% Juvenile Myoclonic Epilepsy (JME), 11% other generalized epilepsies, 15% Mesial Temporal Lobe Epilepsy (mTLE), 54% other partial epilepsies and 9% unclassified epilepsies.

5.2.1.4 Replication Cohort D

The second replication cohort is comprised of an additional 267 multiplex epilepsy families collected by Drs. Samuel Berkovic and Ingrid Scheffer at The University of Melbourne (Victoria, Australia). Approximately 60% of these families have a familial diagnosis of GEFS+ (n=163) and ~40% of families have temporal lobe epilepsy (TLE) patients (the majority of which have >2 individuals, suggestive of non-lesional TLE).

5.2.1.5 Control samples

A healthy control cohort of approximately 2,300 whole-genome (~40%) or exome sequenced (~60%) samples are also currently available in the CHGV; however, only a subset(s) of these samples were available for specific analyses as indicated in the results

section. These control samples were not ascertained for seizure disorders, although we have no immediate way of knowing if some small subset of these samples have unreported epilepsy or have developed epilepsy since the initial collection of these samples. Approximately 77% of this cohort is self-identified as Caucasian. A number of these samples are considered healthy controls with no known disorder (~25%), while the other samples were collected as part of other studies. Neurological disorders with comorbidities with epilepsy have been excluded. Other studies included in this control cohort include amyotrophic lateral sclerosis (~46%), HIV related phenotypes (~17%), kidney disorders (~3.5%), lung disorders (~0.5%), cardiac and vascular disorders (~4%), or other non-neurological disorders (~19%).

A secondary control cohort of ~6,500 subjects was available from the National Heart, Lung, and Blood Institute (NHLBI) Grand Opportunity (GO) Exome Sequencing Project (ESP) for the identification of genes contributing to heart, lung, and blood disorders (Exome Variant Server [EVS], NHLBI ESP, Seattle, WA).

5.2.2 CNV screening

Samples selected for whole-genome or exome sequencing were genotyped on a high-density SNP array and CNV analysis was conducted prior to NGS. Samples were genotyped at CHGV's Genomic Analysis Facility on either the Illumina Human 610-Quad BeadChip (approximately 600,000 SNPs and 60,000 CNV probes) or a Beta-test version of the Illumina human 1M-Duo chip (~800,000 loci). Both the allelic ratios and

SNP intensity data can be extracted from the genotyping results to detect CNVs. The PennCNV algorithm[167] was used to call CNVs based on the B allele frequency (BAF) and the log R ratio (LRR) and basic quality control metrics were applied as described in Heinzen *et al.*[30].

The resulting CNV predictions were then analyzed for evidence of pathogenicity by looking for the presence of: (1) any of the three recurrent CNVs known to increase epilepsy risk: 15q11.2[52], 15q13.3[31], and 16p13.11[52],[30] (2) large heterozygous deletions (>1Mb) which are significantly enriched in epilepsy patients [30] or (3) inclusion of a known epilepsy gene (n=21; [27]) or (4) rare CNVs (absence of CNVs in the same region in samples enrolled in the Database of Genomic Variants, http://dgv.tcag.ca/dgv/app/home). Samples without any evidence of such causal CNVs were then submitted for NGS.

5.2.3 Whole-genome and exome sequencing

5.2.3.1 Sequencing and bioinformatics

Whole-genome or exome sequencing was performed on either the Illumina Genome Analyzer IIx or the Illumina HiSeq2000 platform (Illumina, Inc. San Diego, CA). For exome sequencing, targeted exonic regions and flanking intron-exon boundaries were captured using the SureSelect Human All Exon technology (Aligent Technologies, Santa Clara, CA); the 37.8 Mb and 50 Mb kits were used. The resulting short-sequence reads were aligned to the reference genome (NCBI human genome assembly build 36;

Ensembl core database release 50_361[82]) using the Burrows-Wheeler Alignment (BWA) tool[83] and PCR duplicates were removed using the Picard software (http://picard.sourceforge.net). Genetic differences between each epilepsy genome and the reference genome were identified using the SAMtools variant calling program[84], which identifies both single nucleotide variants (SNVs) and small insertion-deletions (indels). These variants were filtered for high quality variants by requiring: consensus score ≥ 20, quality score ≥ 20, # reads supporting variant ≥ 3 (SNVs) and consensus score ≥ 20, quality score ≥ 50, # reads supporting variant ≥ 3, and ratio of reads supporting variant/reads supporting reference: 0.2-5.0 (indels). SequenceVariantAnalyzer (SVA)[85] software was used to annotate all identified variants and identify variants that were shared by family members and absent or rare in control genomes. The requirements of this study also informed and enabled the development of SVA; particularly, the functions used to filter and list shared variants. These functions were eventually adapted for use in the Analysis Tool for Annotated Variants (ATAV) – a command line statistical software package for performing basic listings and association analysis on annotated variants derived from whole-genome or exome sequencing data (http://redmine2.chgv.lsrc.duke.edu/projects/atav/wiki).

Additionally, the estimation by read depth with single-nucleotide variants (ERDS) algorithm was used to call CNVs in the whole-genome sequenced samples[9].

5.2.3.2 Quality control metrics

Quality control checks were performed on whole-genome and exome sequenced samples. These included: (1) gender check, (2) total number of identified variants, (3) concordance of variants identified by NGS to variants identified with a genotyping chip, (4) overall sequencing coverage, and (5) relatedness between sequenced samples, each detailed below.

Gender check

The gender check was performed by checking the ratio of average coverage across the X and Y chromosomes. The sequencing data confirmed that all NGS samples from the discovery cohorts are the expected gender (Table 20).

Total number of identified variants

The total number of identified variants can indicate samples that are outliers compared the average number of variants observed in other whole-genome or exome samples, suggestive of contaminated samples or other quality control issues. In the discovery cohorts, ~3.6 million SNVs and ~36K indels were identified in the whole-genome sequenced samples (Table 22), consistent with previous reports[1]. In the exome samples (50Mb capture kit) from the discovery cohorts, we found an average of ~680K SNVs and ~2.7K indels.

Concordance

The vast majority of discovery cohort samples were successfully genotyped on either the Illumina Human 610-Quad BeadChip (approximately 600,000 SNPs and 60,000 CNV probes) or a Beta-test version of the Illumina human 1M-Duo chip (~800,000 loci) (Table 20). The SNPs on these arrays provide a means of measuring the concordance between these genotyped SNPs and the variants identified from the whole-genome sequencing data, thus providing an important quality control step. Samples sequenced in the CHGV's genomic analysis facility have an average concordance rate of ~98%[1]. Samples found to have a drastically lower concordance likely indicate a sample swap. One sample, otepi18010m1, had a concordance rate of ~88% (Table 20) and is predicted to have a low level of contamination from another sample (estimated ~19% contamination, data not shown). Due to this finding, we also sequenced a third member of this family and only used otepi18010m1 to check for the presence of variants shared by the two uncontaminated genomes.

Table 20. Sequencing type, prep kit, sequencing platform, gender, ethnicity, genotyping chip, and concordance results for all 69 NGS samples from Discovery Cohorts A and B.

CHGV Family ID	CHGVID	Type	Prep Kit	Platform	Self declared & sequencing inferred gender	XY Average Coverage Ratio	Self Declared Ethnicity	Genotyping Chip	Two Way Concordance Rate
otepia	otepi11221a2	Genome	N/A	GAIIx	F	33.03	White	Human610	99.38%
otepia	ottmn18410	Genome	N/A	GAIIx	F	27.34	White	Human610	99.23%
otepib	ottmn16313	Genome	N/A	GAIIx	M	0.76	Hispanic	Human610	98.93%
otepib	otepi12464b2	Genome	N/A	GAIIx	F	31.8	Hispanic	Human610	99.46%
otepic	otepi14138c1	Genome	N/A	GAIIx	F	32.73	White	Human610	99.49%
otepic	otepi13696c2	Genome	N/A	HiSeq2000	F	33.94	White	Human610	99.61%
otepid	otepi15449d1	Genome	N/A	GAIIx	F	31.77	White	Human610	99.58%
otepid	otepi15075d2	Genome	N/A	GAIIx	F	39.27	White	Human610	99.57%
otepie	otepi14222e1	Genome	N/A	GAIIx	F	30.5	White	Human610	99.45%
otepie	otepi14221e2	Genome	N/A	GAIIx	F	35.78	White	Human610	99.63%
otepif	otepi16923f1	Genome	N/A	GAIIx	F	27.36	White	Human610	99.28%
otepif	otepi17556f2	Genome	N/A	GAIIx	M	0.64	White	Human610	99.52%
otepig	otepi15001g1	Genome	N/A	GAIIx	M	0.65	White	Human610	99.45%
otepig	otepi16934g2	Genome	N/A	GAIIx	F	20.78	White	Human610	99.43%
otepig	otepi15005g3	Genome	N/A	HiSeq2000	F	33.84	White	Human610	99.46%
otepih	otepi18345h1	Genome	N/A	GAIIx	F	33.05	White	Human610	99.25%
otepih	otepi18150h2	Genome	N/A	GAIIx	F	37.6	White	Human610	99.44%
otepi	otepi9778i1	Genome	N/A	GAIIx	F	27.66	White	Human610	99.62%
otepi	otepi743i2	Genome	N/A	GAIIx	M	0.65	White	Human610	99.37%
otepij	otepi10843j1	Exome	37MB	GAIIx	M	0.52	Unknown	Not chipped	N/A
otepij	otepi10694j2	Exome	37MB	GAIIx	M	0.53	Unknown	Not chipped	N/A
otepik	otepi14411k1	Genome	N/A	HiSeq2000	F	35.33	White	Human610	99.49%
otepik	otepi14541k2	Genome	N/A	HiSeq2000	F	33.59	White	Human610	99.42%
otepim	otepi18010m1	Genome	N/A	HiSeq2000	F	34.43	White	Beta	88.19%
otepim	otepi18478m2	Genome	N/A	HiSeq2000	F	32.93	White	Beta	98.24%
otepim	otepi17988m3	Genome	N/A	HiSeq2000	F	32.81	White	Beta	98.51%
otepin	otepi19602n1	Genome	N/A	HiSeq2000	F	33.92	White	Human610	99.39%
otepin	otepi21379n2	Genome	N/A	HiSeq2000	F	33.9	White	Human610	99.15%
otepip	otepi9774p1	Genome	N/A	HiSeq2000	F	22.23	White	Beta	98.41%
otepip	otepi919p2	Genome	N/A	HiSeq2000	M	0.67	White	Beta	98.84%
otepiq	otepi15015q1	Genome	N/A	HiSeq2000	F	33.29	White	Human610	99.52%
otepiq	otepi14992q2	Genome	N/A	HiSeq2000	F	31.26	White	Human610	99.12%
otepir	otepi10478r1	Genome	N/A	HiSeq2000	F	33.19	White	Human610	99.19%
otepir	otepi10818r2	Genome	N/A	HiSeq2000	F	33.82	White	Human610	99.42%
otepis	otepi15880s1	Genome	N/A	HiSeq2000	F	33.43	White	Human610	99.51%
otepis	otepi18097s2	Genome	N/A	HiSeq2000	M	0.61	White	Human610	99.50%
otepit	otepi10425t1	Genome	N/A	HiSeq2000	F	13.65	White	Human610	99.21%
otepiu	otepi7861u1	Genome	N/A	HiSeq2000	M	0.6	White	Human610	99.55%
otepiv	otepi8884v1	Genome	N/A	HiSeq2000	F	36.13	White	Human610	99.50%
otepiw	otepi8089w1	Genome	N/A	HiSeq2000	M	0.74	White	Human610	99.41%
otepix	otepi10554x1	Genome	N/A	HiSeq2000	F	23.78	White	Human610	99.28%
otepiy	otepi7775y1	Genome	N/A	HiSeq2000	M	0.56	White	Human610	99.53%

CHGV Family ID	CHGVID	Type	Prep Kit	Platform	Self declared & sequencing inferred gender	XY Average Coverage Ratio	Self Declared Ethnicity	Genotyping Chip	Two Way Concordance Rate
otepiz	otepi11699z1	Genome	N/A	HiSeq2000	M	0.66	White	Human610	99.47%
otepiaa	otepi13345aa1	Genome	N/A	HiSeq2000	F	19.88	White	Human610	99.48%
otepibb	otepi13220bb1	Genome	N/A	HiSeq2000	M	0.66	White	Human610	99.59%
otepicc	otepi18334cc1	Genome	N/A	HiSeq2000	F	32.34	White	Human610	99.44%
otepidd	otepi14568dd1	Genome	N/A	HiSeq2000	F	33.88	White	Human610	98.24%
otepiee	otepi15721ee1	Genome	N/A	HiSeq2000	F	35.81	Hispanic	Human610	99.52%
bkva	bkv5675a1	Exome	50MB	HiSeq2000	M	0.44	White	Human610	99.68%
bkva	bkv5912a2	Exome	50MB	GAIIx	M	0.44	White	Not chipped	N/A
bkva	bkv2343a3	Exome	50MB	HiSeq2000	M	0.47	White	Human610	99.76%
bkvb	bkv16243b1	Exome	50MB	HiSeq2000	F	69.59	White	Human610	99.75%
bkvb	bkv2347b2	Exome	50MB	HiSeq2000	F	74.7	White	Human610	99.77%
bkvc	bkv1299c1	Exome	50MB	GAIIx	F	94.18	White	Human610	99.77%
bkvc	bkv1300c2	Exome	50MB	HiSeq2000	F	63.22	White	Human610	99.71%
bkvd	bkv6443d1	Exome	50MB	HiSeq2000	M	0.47	White	Human610	99.66%
bkvd	bkv7632d2	Exome	50MB	HiSeq2000	M	0.48	White	Human610	99.61%
bkve	bkv187e1	Exome	50MB	HiSeq2000	M	0.48	White	Human610	99.74%
bkve	bkv38e2	Exome	50MB	HiSeq2000	M	0.59	White	Human610	99.79%
bkvf	bkv6622f1	Exome	50MB	HiSeq2000	F	64.85	White	Human610	99.77%
bkvf	bkv6648f2	Exome	50MB	HiSeq2000	F	37.01	White	Human610	99.77%
bkvg	bkv715g1	Exome	50MB	HiSeq2000	M	0.46	White	Human610	99.67%
bkvg	bkv777g2	Exome	50MB	HiSeq2000	M	0.47	White	Human610	99.71%
bkvh	bkv19260h1	Exome	50MB	HiSeq2000	M	0.48	White	Human610	99.71%
bkvh	bkv19263h2	Exome	50MB	HiSeq2000	M	0.46	White	Human610	99.68%
bkvi	bkv8227i1	Exome	50MB	HiSeq2000	M	0.47	White	Human610	99.69%
bkvi	bkv17151i2	Exome	50MB	HiSeq2000	M	0.48	White	Human610	99.70%
bkvj	bkv7065j1	Exome	50MB	HiSeq2000	F	77.5	White	Human610	99.79%
bkvj	bkv7068j2	Exome	50MB	HiSeq2000	F	63.83	White	Human610	99.79%

Overall sequencing coverage

Previously, our lab established that once the overall coverage exceeds 30–35x, the concordance rate between sequencing and genotyping will be greater than ~98.6% indicating this coverage threshold is important for accurate variant calling[1]. For all discovery cohort samples a base was considered covered if ≥5 reads spanned this base. On average, the whole-genome sequenced samples had a read depth of 34.25x and the exome sequenced samples had a read depth of 57.91x (Table 22).

Relatedness (IBD calculation)

An estimate of the relationship between familial samples was obtained by using sequencing data to estimate identity by descent (IBD) and comparing this value to the expected IBD value based on the known relationship between the sequenced individuals. IBD was estimated in the following way: let x_{ij} be rare variations in genomes *i* and *j* (shared variants). Let x_i be the number of rare variations in genome *i* (includes variants overlapping with *j*). Let x_j be rare variations in genome *j* (includes variants overlapping with *i*). Then, $x_{ij} = \frac{x_i + x_j}{2} \times p$, where *p* = probability that a variant present in *i* is also present in *j*. When the included variants are sufficiently rare, this is ~IBD. In solving for *p*, IBD can be calculated using the following equation: $p = \frac{2x_{ij}}{x_i + x_j}$.

We note that the IBD estimates obtained from the sequencing data tend to be systematically lower than expected (Figure 31). This is a trend we have observed across all familial data sets, not just this cohort. It is plausible that the numbers may still differ slightly because of false positive variant calls and/or use of an improper set of control genomes for assessing variant frequencies. Importantly, we are fairly close to the expected IBD, which indicates that we are likely capturing the majority of shared variants in these families.

In the discovery cohorts, the estimated IBD values are close enough to the expected values that there is a low likelihood of a sample swap confirming sample identity (Table 21, Figure 31).

Table 21. IBD values estimated from NGS samples.

CHGV Family ID	Average exonic read depth across both samples	Proband	Other relative	Relationship of other relative to sequenced proband	Expected IBD	Estimated IBD	Number of controls used to assess rarity of variants
otepia	32.11	otepi11221a2	ottmn18410	Aunt-Uncle	0.250	0.169	266
otepib	27.99	ottmn16313	otepi12464b2	First cousin once removed	0.063	0.089	266
otepic	37.80	otepi14138c1	otepi13696c2	First cousin	0.125	0.117	266
otepid	57.65	otepi15449d1	otepi15075d2	First cousin once removed	0.063	0.054	266
otepie	71.36	otepi14222e1	otepi14221e2	Second cousin	0.031	0.020	266
otepif	31.56	otepi16923f1	otepi17556f2	First cousin once removed	0.063	0.090	266
otepig	28.72	otepi15001g1	otepi16934g2	Great niece-nephew	0.063	0.047	266
otepih	28.05	otepi18345h1	otepi18150h2	First cousin	0.125	0.102	266
otepii	32.11	otepi9778i1	otepi743i2	Aunt-Uncle	0.250	0.194	266
otepij	96.19	otepi10843j1	otepi10694j2	First cousin once removed	0.063	0.122	255
otepik	27.90	otepi14411k1	otepi14541k2	Half first cousin	0.063	0.060	385
otepim	33.20	otepi18010m1	otepi18478m2	Aunt-Uncle	0.250	0.178	385
otepim	28.70	otepi18010m1	otepi17988m3	Sibling	0.250	0.178	385
otepin	28.19	otepi19602n1	otepi21379n2	Grandparent	0.250	0.148	385
otepip	32.90	otepi9774p1	otepi919p2	First cousin	0.125	0.115	385
otepiq	31.83	otepi15015q1	otepi14992q2	Sibling	0.500	0.452	385
otepir	28.69	otepi10478r1	otepi10818r2	Parent	0.500	0.395	385
otepis	34.55	otepi15880s1	otepi18097s2	First cousin	0.125	0.125	385
bkva	56.79	bkv5675a1	bkv5912a2	First cousin	0.125	0.072	371
bkva	52.24	bkv5675a1	bkv2343a3	Second cousin	0.125	0.072	371
bkvb	56.35	bkv16243b1	bkv2347b2	Aunt-Uncle	0.250	0.184	371
bkvc	56.82	bkv1299c1	bkv1300c2	Aunt-Uncle	0.250	0.202	371
bkvd	52.87	bkv6443d1	bkv7632d2	Aunt-Uncle	0.250	0.247	371
bkve	58.75	bkv187e1	bkv38e2	First cousin	0.125	0.119	371
bkvf	64.74	bkv6622f1	bkv6648f2	First cousin	0.125	0.125	371
bkvg	47.49	bkv715g1	bkv777g2	Third cousin once removed	0.016	0.017	371
bkvh	47.32	bkv19260h1	bkv19263h2	First cousin once removed	0.063	0.025	371
bkvi	46.11	bkv8227i1	bkv17151i2	Niece-Nephew	0.250	0.164	371
bkvj	52.60	bkv7065j1	bkv7068j2	Child	0.500	0.377	371

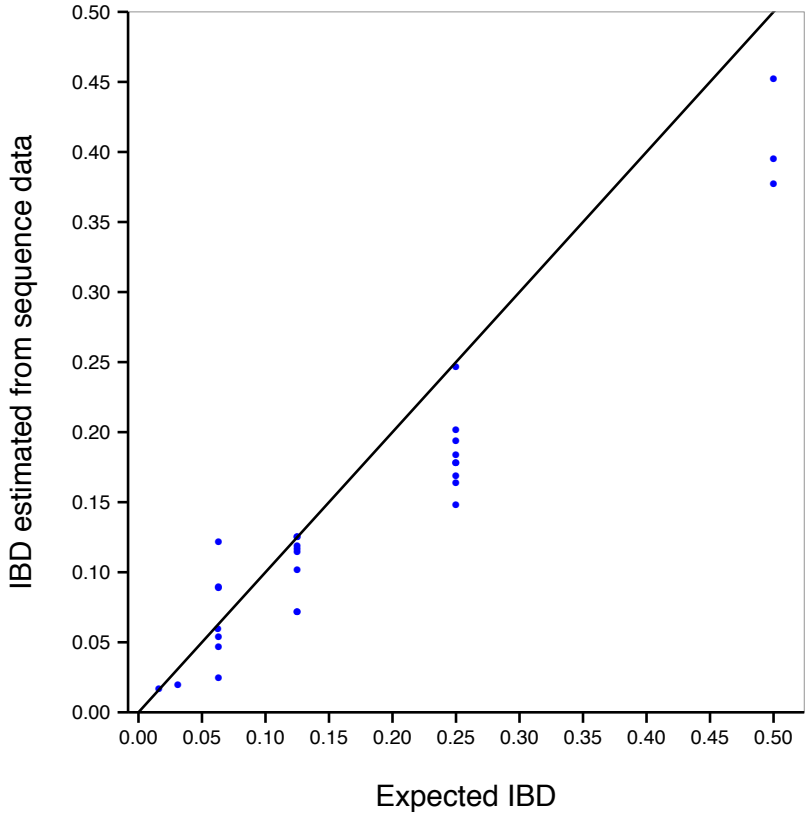


Figure 31. Expected IBD vs. IBD estimated from NGS samples.
Estimated IBD values from NGS samples in Table 21 (blue dots) and their expected IBD based on relatedness (line).

Table 22. Variant counts, transition to transversion ratio (TiTv), dbSNP overlap, and coverage results for all 69 NGS samples from Discovery Cohorts A and B.

For whole-genome sequenced samples, covered bases (≥5x), average read depth, and % covered bases values are based only on autosomal exonic bases. *The data for sample bkv38e2 was lost in an attempt to convert to Genome Reference Consortium Human Genome build 37.

CHGV Family ID	CHGVID	SeqType	Number of SNVs passing QC	Number of indels passing QC	TiTv	dbSNP Overlap	Genome wide covered bases (≥1x)	Genome wide average read depth	Covered exonic bases (≥5X)	Average read depth (exonic)	Total % covered exonic bases (5x)
otepia	otepi11221a2	Genome	3,647,150	662,821	1.99	94.45	2,831,011,250	39.55	62,523,558	31.77	95.50%
otepia	ottmn18410	Genome	3,574,779	603,343	2.04	97.05	2,834,010,780	32.26	64,680,869	32.45	98.79%
otepib	ottmn16313	Genome	3,738,998	598,898	2.04	96.32	2,852,646,236	27.18	64,622,306	28.42	98.70%
otepib	otepi12464b2	Genome	3,915,236	832,535	2.00	94.45	2,830,921,952	33.5	62,579,260	27.55	95.58%
otepic	otepi14138c1	Genome	3,655,012	695,434	2.02	95.46	2,832,108,608	31.34	63,650,458	26.55	97.22%
otepic	otepi13696c2	Genome	3,719,501	699,253	1.96	93.27	2,833,599,838	56.49	64,640,764	49.04	98.73%
otepid	otepi15449d1	Genome	3,645,717	684,065	1.99	94.28	2,831,798,486	44.22	63,548,328	37.08	97.06%
otepid	otepi15075d2	Genome	3,659,115	888,041	1.98	92.98	2,832,620,322	41.63	64,460,322	78.21	98.46%
otepie	otepi14222e1	Genome	3,628,631	805,918	2.01	94.91	2,831,826,162	40.46	63,684,138	77.05	97.27%
otepie	otepi14221e2	Genome	3,664,181	833,465	2.02	95.34	2,832,968,725	31.94	64,417,389	65.66	98.39%
otepif	otepi16923f1	Genome	3,594,850	691,921	2.02	95.75	2,832,832,496	37.95	62,388,552	30.7	95.29%
otepif	otepi17556f2	Genome	3,609,789	687,467	2.02	94.64	2,851,628,888	37.96	63,440,191	32.41	96.90%
otepig	otepi15001g1	Genome	3,635,524	658,888	2.02	95.43	2,850,637,335	26.77	64,219,415	28.46	98.09%
otepig	otepi16934g2	Genome	3,630,479	673,152	2.03	95.91	2,835,781,477	31.88	63,367,894	28.98	96.79%
otepig	otepi15005g3	Genome	3,651,292	663,257	2.00	94.58	2,830,470,161	41.93	62,127,057	33.21	94.89%
otepih	otepi18345h1	Genome	3,628,439	691,190	2.02	95.32	2,831,745,685	33.09	62,969,484	27.1	96.18%
otepih	otepi18150h2	Genome	3,623,588	673,688	2.03	95.98	2,830,956,588	33.35	63,190,458	29	96.52%
otepi	otepi9778i1	Genome	3,671,785	707,778	2.01	95.18	2,835,646,283	38.54	64,383,387	34	98.34%
otepi	otepi743i2	Genome	3,640,259	681,110	2.00	95.42	2,851,369,814	32.91	63,740,767	30.21	97.36%
otepij	otepi10843j1	Exome	36,336	2,444	2.88	98.43	N/A	N/A	36,623,435	97.49	95.50%
otepij	otepi10694j2	Exome	35,456	2,436	2.84	98.53	N/A	N/A	36,825,236	94.89	96.20%
otepik	otepi14411k1	Genome	3,671,098	679,737	2.02	95.42	2,830,930,704	33.98	62,699,084	27.58	95.77%
otepik	otepi14541k2	Genome	3,680,517	683,913	1.98	94.26	2,831,703,148	34.06	63,043,727	28.21	96.29%
otepim	otepi18010m1	Genome	3,854,021	664,033	2.02	95.8	2,832,684,334	41.37	63,356,342	34.61	96.77%

CHGV Family ID	CHGVID	SeqType	Number of SNVs passing QC	Number of indels passing QC	TiTv	dbSNP Overlap	Genome wide covered bases (≥1x)	Genome wide average read depth	Covered exonic bases (≥5X)	Average read depth (exonic)	Total % covered exonic bases (5x)
otepim	otepi18478m2	Genome	3,602,777	637,798	2.01	95.16	2,830,292,911	39.84	61,968,573	31.78	94.65%
otepim	otepi17988m3	Genome	3,631,632	643,617	2.03	95.87	2,830,861,142	29.79	61,812,816	22.78	94.41%
otepin	otepi19602n1	Genome	3,666,179	662,978	2.02	94.9	2,831,583,110	37.26	62,459,955	30.21	95.40%
otepin	otepi21379n2	Genome	3,588,936	636,116	2.01	95.46	2,830,311,497	32.75	61,292,225	26.17	93.62%
otepip	otepi9774p1	Genome	3,708,417	666,527	1.98	93.76	2,837,113,711	40.18	62,665,712	32.52	95.72%
otepip	otepi919p2	Genome	3,749,585	676,111	1.99	94.15	2,852,297,905	36.96	63,887,233	33.27	97.58%
otepiq	otepi15015q1	Genome	3,667,565	665,060	2.00	95.27	2,831,984,577	36.85	63,088,699	30.6	96.36%
otepiq	otepi14992q2	Genome	3,659,215	662,135	2.01	95.37	2,831,650,948	40.25	62,884,721	33.06	96.05%
otepir	otepi10478r1	Genome	3,652,837	651,398	1.99	94.1	2,831,310,917	32.77	62,436,583	26.92	95.37%
otepir	otepi10818r2	Genome	3,651,855	655,863	1.98	93.98	2,830,993,449	37.93	62,270,342	30.45	95.11%
otepis	otepi15880s1	Genome	3,649,015	666,807	2.02	95.31	2,832,238,734	42.59	63,329,691	35.03	96.73%
otepis	otepi18097s2	Genome	3,637,681	661,919	2.00	94.53	2,851,518,116	40.17	63,148,885	34.06	96.45%
otepit	otepi10425t1	Genome	3,705,301	685,331	1.99	94.92	2,843,316,924	37.97	64,010,561	33.11	97.77%
otepiu	otepi7861u1	Genome	3,655,327	667,318	1.99	94.94	2,852,137,675	41.39	63,662,704	35.86	97.24%
otepiv	otepi8884v1	Genome	3,726,802	668,044	2.00	94.24	2,832,023,764	39.04	63,157,617	32.34	96.47%
otepiw	otepi8089w1	Genome	3,712,556	679,454	1.99	93.81	2,850,780,724	42.07	62,570,208	34.66	95.57%
otepix	otepi10554x1	Genome	4,057,788	722,080	2.02	94.91	2,835,423,639	40.24	62,115,344	32.08	94.87%
otepiy	otepi7775y1	Genome	3,664,915	670,296	1.96	93.58	2,851,224,467	39.49	63,191,960	33.63	96.52%
otepiz	otepi11699z1	Genome	3,627,107	651,996	2.01	94.66	2,850,941,642	39.83	62,748,765	33.04	95.84%
otepiaa	otepi13345aa1	Genome	3,653,602	666,476	1.99	94.32	2,838,280,879	35.16	63,315,384	29.38	96.71%
otepibb	otepi13220bb1	Genome	3,639,471	657,538	2.02	95.47	2,851,427,988	40.11	63,323,293	34.16	96.72%
otepicc	otepi18334cc1	Genome	3,643,449	649,554	2.01	94.83	2,831,862,516	41.57	62,215,816	32.89	95.03%
otepidd	otepi14568dd1	Genome	3,594,686	654,699	2.02	95.95	2,830,812,096	33.97	61,465,925	26.56	93.88%
otepiee	otepi15721ee1	Genome	3,726,866	673,988	2.01	95.09	2,831,579,624	39.78	62,912,817	32.52	96.09%
bkva	bkv5675a1	Exome	34,811	2,623	2.72	98.23	N/A	N/A	45,995,153	44.58	91.30%
bkva	bkv5912a2	Exome	35,545	2,680	2.75	99.08	N/A	N/A	46,838,420	68.99	92.90%
bkva	bkv2343a3	Exome	37,563	2,868	2.73	98.87	N/A	N/A	47,785,934	59.9	94.80%
bkvb	bkv16243b1	Exome	36,937	2,838	2.71	98.28	N/A	N/A	47,320,035	55.6	93.90%
bkvb	bkv2347b2	Exome	37,040	2,864	2.67	98.03	N/A	N/A	47,482,960	57.09	94.20%
bkvc	bkv1299c1	Exome	36,832	2,783	2.77	99.15	N/A	N/A	47,730,777	62.92	94.70%

CHGV Family ID	CHGVID	SeqType	Number of SNVs passing QC	Number of indels passing QC	TiTv	dbSNP Overlap	Genome wide covered bases (≥1x)	Genome wide average read depth	Covered exonic bases (≥5X)	Average read depth (exonic)	Total % covered exonic bases (5x)
bkvc	bkv1300c2	Exome	36,266	2,753	2.77	98.87	N/A	N/A	46,830,741	50.72	92.90%
bkvd	bkv6443d1	Exome	35,947	2,701	2.74	98.81	N/A	N/A	46,989,031	57.34	93.20%
bkvd	bkv7632d2	Exome	36,200	2,693	2.71	98.31	N/A	N/A	46,969,942	48.4	93.20%
bkve	bkv187e1	Exome	36,212	2,754	2.78	98.84	N/A	N/A	47,233,065	64.84	93.70%
bkve	bkv38e2	Exome	35,668	2,730	*	*	N/A	N/A	47,083,340	52.65	93.40%
bkvf	bkv6622f1	Exome	37,331	2,855	2.75	98.89	N/A	N/A	48,017,583	87.04	95.30%
bkvf	bkv6648f2	Exome	35,442	2,710	2.68	97.92	N/A	N/A	46,228,345	42.44	91.70%
bkvg	bkv715g1	Exome	34,790	2,672	2.74	98.54	N/A	N/A	46,336,880	44.27	91.90%
bkvg	bkv777g2	Exome	35,408	2,716	2.71	98.88	N/A	N/A	46,886,201	50.71	93.00%
bkvh	bkv19260h1	Exome	35,024	2,577	2.74	98.67	N/A	N/A	46,675,447	51.69	92.60%
bkvh	bkv19263h2	Exome	35,600	2,808	2.77	98.74	N/A	N/A	46,536,999	42.95	92.30%
bkvi	bkv8227i1	Exome	36,980	2,806	2.70	97.93	N/A	N/A	46,910,734	50.56	93.10%
bkvi	bkv17151i2	Exome	35,744	2,735	2.71	98.34	N/A	N/A	46,375,132	41.65	92.00%
bkvj	bkv7065j1	Exome	36,084	2,782	2.77	99.01	N/A	N/A	47,386,584	55.73	94.00%
bkvj	bkv7068j2	Exome	36,141	2,781	2.73	98.41	N/A	N/A	46,979,440	49.47	93.20%

5.2.4 Custom genotyping

5.2.4.1 Variants and samples

Candidate variants were genotyped using a custom designed iSelect genotyping chip (Illumina). The candidate variants from this project were combined with variants of interest from several other projects from our lab [56,168] and collectively, approximately 12,000 variants were genotyped across 7,000 samples. For the familial epilepsy project, a total of 744 candidate variants were selected for custom genotyping; only 609 of these variants passed the initial design phase and subsequent QC filtering. These variants were genotyped in a total of 949 unrelated epilepsy samples from Replication Cohort C (the majority were generalized epilepsy samples) and 1,818 unrelated neurologically normal controls. As a positive control, a single proband from each of the twenty discovery families was also included (Discovery Cohorts A (otepia-otepij) and B (bkva-bkvj)).

Additionally, for nine of the families from Discovery Cohort A (a-i), all relatives with available DNA (both affected and unaffected) were also genotyped to assess cosegregation of candidate variants from these families.

5.2.4.2 Quality control

A gender comparison was performed using custom chip determined genders and self-reported genders; samples with mismatched genders were excluded from further analysis. Samples with a call rate below 0.95 were excluded from further

analysis. For the subset of samples previously genotyped on a non-custom Illumina genome-wide genotyping chip, we compared the genotyping calls of the 170 variants genotyped on both platforms and removed samples with <95% concordance. Additionally, variants were subjected to quality control measures. For the subset of samples with whole-genome or exome sequencing data, a concordance check was performed; variants with discordant calls between the genotyping and sequencing data (that could not be reconciled with visual inspection of the genotyping or sequencing data) were excluded from further analysis.

5.2.4.3 Analysis

In order to assess the resulting genotype data in a high-throughput fashion, I generated a custom R script to tally genotype counts across user-defined phenotypes of interest. Case-control status was assigned to run an allelic Fisher's exact test (FET) or cosegregation was assessed by assigning affectation status (including possible carrier status (e.g., unaffected parents of affected individuals)) and then tallying carriers to noncarriers.

To formally assess enrichment of variants in epilepsy cases vs. controls, PLINK [89] was used to perform a logistic regression. Only a single affected individual was included from each family and the sporadic cohort of unrelated epilepsy cases and controls were included. To adjust for population stratification, the EIGENSTRAT program[169] was used to create covariates prior to running the logistic regression.

Eigenvectors were created from genome-wide chip data and thus we used only the subset of samples that were (i) Caucasian and (ii) genotyped on a genome-wide genotyping chip. Thus the tested sample set (after eliminating any with poor genotyping data from the GWAS chip) was limited to 520 cases and 869 controls (n=1389). The resulting population substructure-corrected logistic regression p-values were plotted using a Quantile-quantile plot (Q-Q plot). Variants found in less than six samples were not plotted. A Q-Q plot is a diagnostic plot that compares the distribution of observed test statistics with the distribution expected under the null hypothesis.

5.2.5 Cosegregation testing

Cosegregation testing was completed using a combination of methods including Sanger sequencing, TaqMan genotyping, and custom genotyping chips (iSelect). Enrichment testing in case-control cohorts was accomplished using the later two methods.

5.2.6 Prioritizing candidate variants from sequenced epilepsy genomes (PVEG)

A score was developed for prioritizing candidate variants with the aim that genetic variants that are likely to cause epilepsy are readily recognizable amongst all identified variants in NGS studies. The Prioritization of candidate variants from sequenced epilepsy genomes (PVEG) score uses known and phenotypically relevant information about each **gene** harboring a qualifying variant. This gene-based score relies heavily on

existing databases and key publications with information about genes that are relevant to epilepsy (Table 23).

The gene-based score is additive, with one point given for each unique category that classifies a given gene. All gene name scores were converted to HGNC names (http://www.genenames.org) and then annotation and score calculation was performed in a custom R script. When looking at all 19,092 HGNC genes, 1,760 genes (~9%) have a PVEG score > 0. The highest scoring PVEG gene is *SCN1A*, with a score of 12.

Table 23. The resources used to calculate PVEG genic scores.

The "gene count" is not necessarily unique (some genes are in multiple categories). References for these resources: known epilepsy genes [27]; 15q11.2 microdeletion [52]; 15q13.3 microdeletion [31,170]; 15q13.3 microdeletion [31,170]; 16p13.11 microdeletion [30,52]; Genes within epilepsy associated >1MB deletions [30]; HGMD (http://www.hgmd.cf.ac.uk/ac/index.php); CarpeDB (http://www.carpedb.ua.edu); Mouse model epilepsy susceptibility genes [64] (one gene was added to this category based on personal communication with the author); EpiGAD (http://www.epigad.org); Genetics Home Reference (http://ghr.nlm.nih.gov); Genes harboring a rare variant (absent in controls) in GGE cases[56]; Genes harboring *de novo* mutations in autism patients [58,59,62,63]; Subset of ion channel genes [65]; finally, the last categories are all based on Lemke *et al.* [43].

PVEG category	Database access date or publication year	Gene count
Known epilepsy genes	2010	21
Genes within the 15q11.2 microdeletion	2010	27
Genes within the 15q13.3 microdeletion	2009	15
Genes within the 16p13.11 microdeletion	2010	16
Genes within epilepsy associated >1MB deletions	2010	373
Human Gene Mutation Database (HGMD) (phenotype = epilepsy)	January 12th, 2012	90
Carpe DB epilepsy genes	January 12th, 2012	204
Mouse model epilepsy susceptibility genes (homologous genes)	2009	104
EpiGAD susceptibility genes	January 12th, 2012	111
Genetics Home Reference: conditions with seizures	February 20th, 2012	316
Genetics Home Reference: genes associated with seizures	February 20th, 2012	128
Genes harboring a rare variant (absent in controls) in GGE cases	2012	157
Genes harboring *de novo* mutations in autism patients	2012	578
Subset of ion channel genes	2011	237
Congenital disorders of glycolization (Lemke)	2012	23
Disorders of peroxisome biogenesis (Lemke)	2012	9
Disorders of the Ras-MAPK pathway (Lemke)	2012	14
Early infantile epileptic encephalopathies (Lemke)	2012	30
Epilepsy in X-linked mental retardation (Lemke)	2012	26
Generalized-myoclonic-absence epilepsies febrile seizures (Lemke)	2012	37
Holoprosencephaly (Lemke)	2012	8
Hyperekplexia (Lemke)	2012	5
Joubert syndrome and related disorders (Lemke)	2012	10
Leukodystrophies (Lemke)	2012	20
Lysosomal storage disorders (Lemke)	2012	29
Migraine (Lemke)	2012	6
Neuronal ceroid lipofuscinosis (Lemke)	2012	8
Neuronal migration disorders (Lemke)	2012	31
Selected inborn errors of metabolism (Lemke)	2012	32
Selected mitochondrial disorders (Lemke)	2012	35
Severe microcephaly-pontocerebellar hypoplasia (Lemke)	2012	22
Syndromic disorders with epilepsy and others (Lemke)	2012	29

5.2.7 Custom capture and sequencing

5.2.7.1 Sequencing and bioinformatics

HaloPlex technology (Agilent, Santa Clara, CA) was used to target a subset of candidate epilepsy genes. This technology was selected because, at the time, it offered capture of 500Kb of sequence across hundreds of samples in at relatively low cost. HaloPlex uses restriction enzyme fragmentation to digest the target DNA.

An adapted version of the CHGV's NGS pipeline was used for this custom capture. Adapter sequences were removed from the reads using cutadapt (http://code.google.com/p/cutadapt/). Read one and read two were independently aligned to the reference genome (Genome Reference Consortium Human Genome build 37 (GRCh37)) using BWA. Due to the use of restriction enzyme digestion in the HaloPlex protocol, PCR duplicates were **not** removed. After alignment, GATK ClipReads was applied to mask the first five bases of all reads. These bases are known to give incorrect allele ratios due to a drop in the hybridization/ligation/digestion efficiency of the HaloPlex protocol.

Variants were called using the Genome Analysis Toolkit (GATK) [86]. GATK's Unified Genotyper was run with two modifications: the haplotype score filter was removed and the mapping quality filter was raised to remove reads with MQ < 35.0. Indel realignment and base quality score recalibration were applied. Variants were then annotated using SnpEff (http://snpeff.sourceforge.net/). Variant analysis was limited to those falling in the restricted region, which is the intersection between the covered

regions (genomic regions expected to be sequenced based on this custom design for target enrichment; 1171.56kb) and targeted regions (used to select the probes; 486.69kb).

One sample (bkv6464ab1) was sequenced after preparation with the HaloPlex custom kit and an exome prep kit (Nimblegen SeqCap EZ V3); this was critical for establishing the bioinformatics pipeline used for all HaloPlex samples.

5.2.7.2 Coverage

The targeted regions comprise ~500 Kb of the genome, however the HaloPlex technology required that we sequence ~1.2Mb of the genome. After sequencing the first plate of samples prepared with the custom HaloPlex on a single flow cell (96 samples multiplexed) and analyzing the resulting coverage data (data not shown), we decided we needed to add a second flow cell to obtain sufficient coverage across the targeted regions. Sequencing samples across two flow cells resulted in ~1000x average coverage across the restricted regions. If we define a covered base as ≥3 reads (3x) spanning this site, then ~95% of the restricted region bases are covered (Figure 32). However, 3x is inadequate for accurate variant calling. The coverage decreases at 10x and 20x (Figure 32); only ~89.5% of the restriction region bases are covered at 20x, which is reflective of uneven coverage across the restricted regions. There is nothing "abnormal" about the poorly sequenced regions in terms of GC-content or repetitive elements (data not shown) and these same regions are captured using different sample preparation

methods, indicating that the unevenness of the HaloPlex coverage is inherent to the technology and not the design/selection of the target regions themselves.

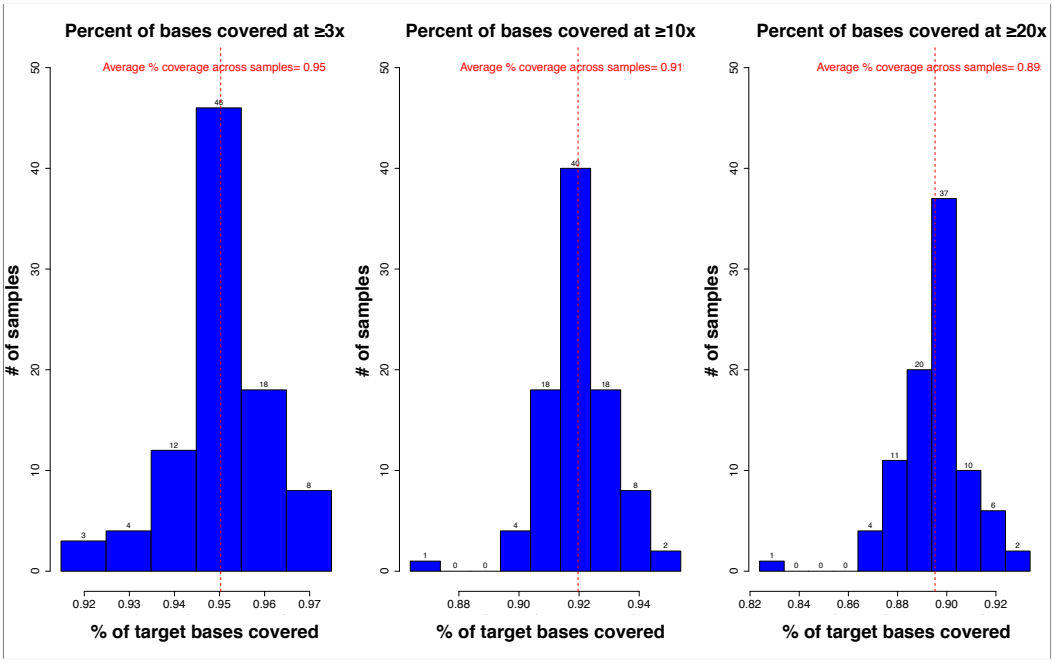


Figure 32. HaloPlex sequencing coverage of targeted regions at 3x, 10x, and 20x in 91 samples from plate 1.

Using the bkv6464ab1 sample to compare coverage resulting from HaloPlex and exome sequencing, we found ~89.3% of the restricted bases covered (20x) in the HaloPlex sample and 94.1% covered in the exome-sequenced version of this sample. For the 2,581 targeted "regions", 56% were covered equally well (or equally poorly) in both samples. For the 1,125 discordant regions, HaloPlex coverage was better than exome coverage for 356 regions (31.6%)(Figure 33).

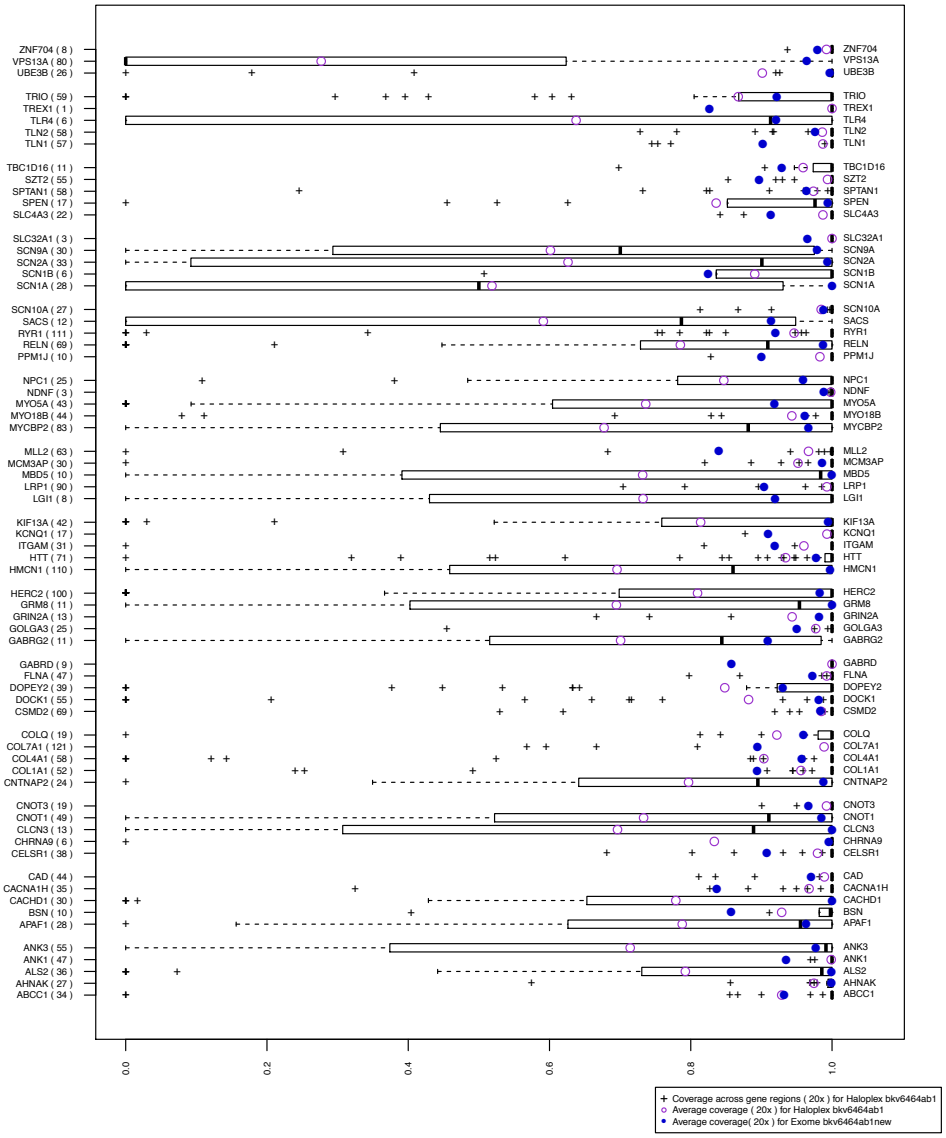


Figure 33. Coverage (20x) across the 68 resequenced genes as sequenced after exome or HaloPlex sample prep.

The coverage for each target region (usually an exon or part of an exon) across each gene for HaloPlex sample bkv6464ab1. The number of target regions are listed in parentheses after gene name. A value of 0.0 means no coverage and a value of 1.0 means 100% coverage. The purple open dots represent the average coverage for the entire gene in the HaloPlex sample; the blue filled dots represent the average coverage for the entire gene in the exome sample. The average coverage for exome is often (but not always) better than HaloPlex and coverage for certain genes are dramatically improved in the exome sample (e.g., *SCN9A* and *SCN2A*).

5.2.7.3 Additional variant filtering

Visual inspection (Integrative Genomics Viewer (IGV)) of variants present exclusively in the HaloPlex bkv6464ab1 sample and absent from the exome bkv6464ab1 sample revealed artifacts introduced by the HaloPlex technology that were not corrected in the bioinformatics pipeline. The rare allele for these variants was consistently found only near the end of reads, suggestive of an error. Rather than adjusting the original bioinformatics pipeline and rerunning all ~470 samples, we used the read position rank sum (RPRS) test. The read position rank sum (RPRS) value is provided for all heterozygous variants called by GATK and provides a score based on the distance from the end of the read for reads with the alternate allele. After establishing that the RPRS values for the HaloPlex variants did not come from a normal distribution, we used a Gaussian mixture model and determined that an RPRS threshold of >-2.9 would eliminate the majority of artifacts while retaining high quality variants (data not shown). To directly evaluate the proposed RPRS threshold, two populations of variants were visually inspected: singleton variants (present in only one sample) and nonsingleton variants (present in multiple samples). A total of 80 singleton variants and 60 nonsingletons (249 variant-carrier pairs) were visually inspected and classified as true positive, likely true positive, likely false positive, or false positive using the rules outlined in Table 24. These variants were selected at random from three RPRS bins

surrounding the proposed threshold (approximately: less than -2.2, equal to -2.9, and greater than -4.0).

Table 24. Rules for classification of variants as true or false positive after visual inspection in IGV.

Inspection Rules	Classification
50:50 reads (var:ref) and high coverage	**True positive**
Middle of read but low coverage	Likely true positive
Not at extreme end (last 5bp) but near ends and ~25% of reads/low coverage	Likely true positive
Variant allele found exclusively at the extreme end of read (and less than ~25% of reads)	**False positive**
Variant allele calls seen on a mixture of extreme ends of reads and nonextreme ends, but majority of reads support the reference allele with high coverage	Likely false positive
Not at extreme end (last 5bp) but near ends and less than 25% of reads	Likely false positive
Within last 10bp of the reads and var:ref ratio is roughly 50:50 but low overall coverage and a rare haplotype at ends (usually ~5 variants in this region)	Likely false positive

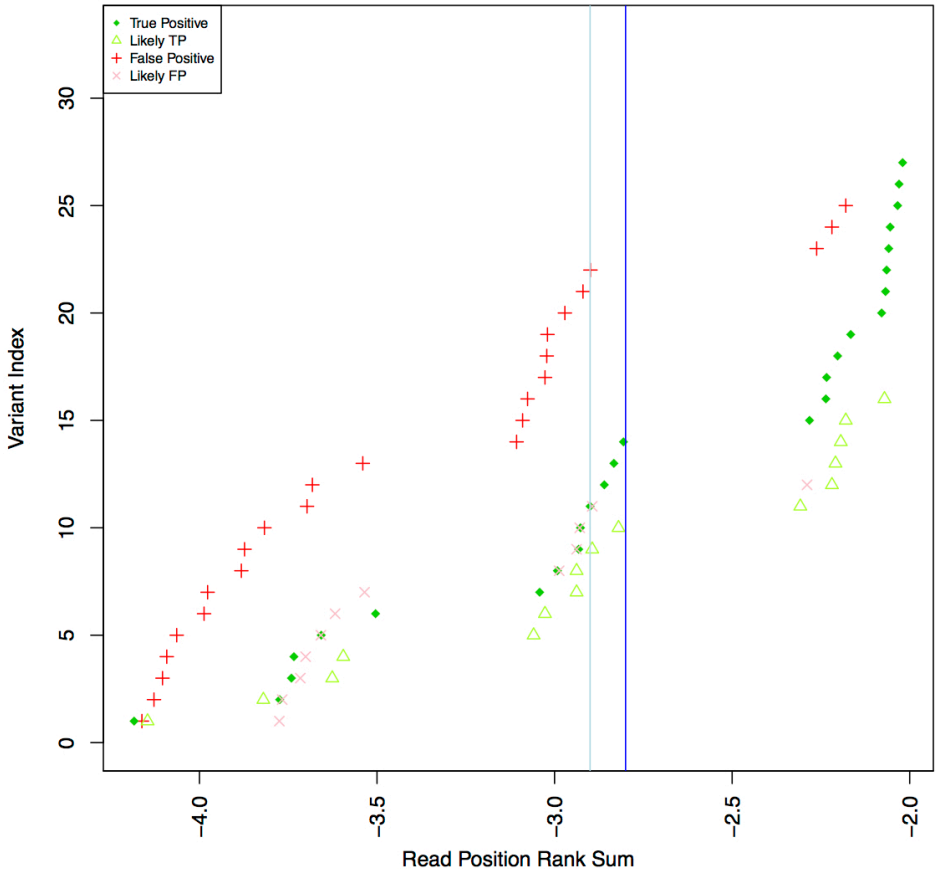


Figure 34. Classification of 80 singleton variants by Read Position Rank Sum

Using a threshold of RPRS > -2.9 (light blue line, Figure 34) successfully removed the majority of the false positive variant calls and retained the majority of true variants (since the vast majority of true ones are much higher than –2.9). We then compared the distribution of true and false positives by looking at ten additional QC metrics to determine if applying an additional metric could further distinguish true from false positive variants. Use of the Quality score accomplished just this with true positive variants always having a quality score of >1000 (Figure 35).

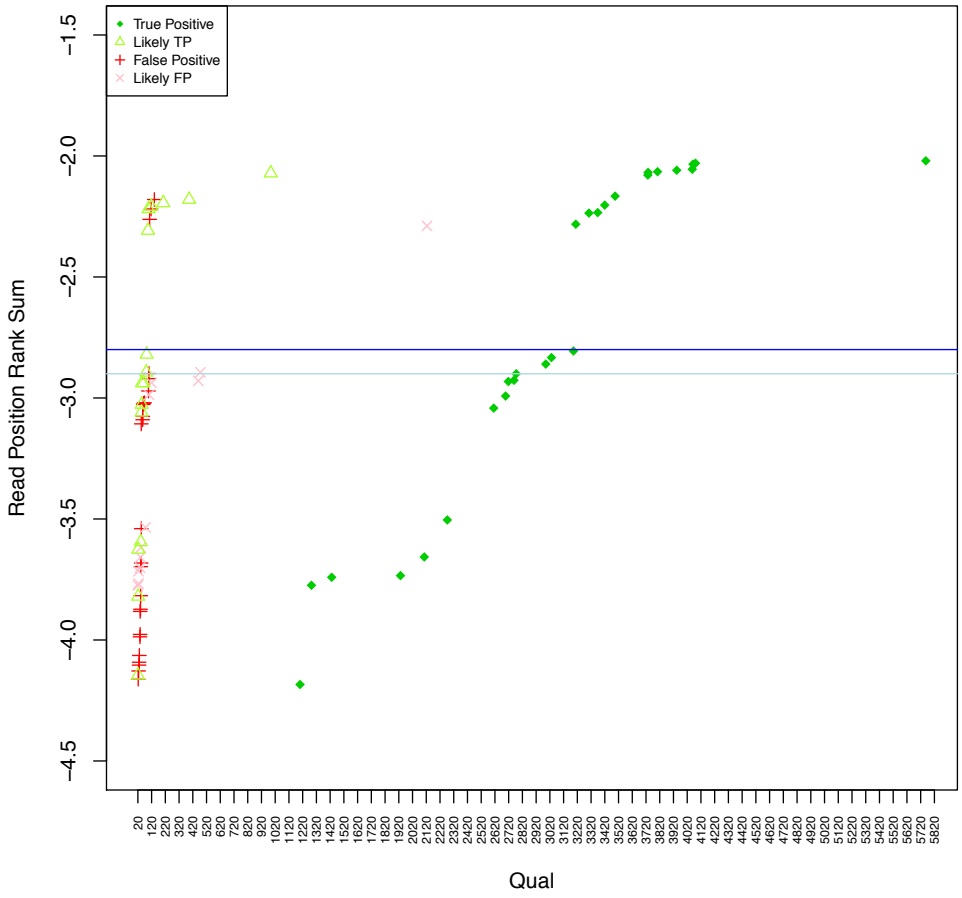


Figure 35. Classification of 80 singleton variants by Read Position Rank Sum and Quality

Inspection of nonsingleton variants demonstrated that the RPRS values vary dramatically between carriers of the same variant but variants were all true positives or all false positives, regardless of the QC values in individual carriers (Appendix C). The rules for variants are applied such that a nonsingleton variant is either included or excluded across all carriers. The three rules, applied sequentially, were: (i) remove all variants with a minimum RPRS value of ≤ -10.0, (ii) otherwise, keep all variants with a

maximum RPRS value of ≥ -2.9, (iii) then, remove variants where all carriers have a variant Qual < 1,000 (unless maxRPRS is ≥ -2.9). In total, this process eliminated ~16% of all unique identified variants across the 473 HaloPlex sequenced samples.

5.2.7.4 Analysis

All variant listings and gene-based collapsing methods were applied using the Analysis Tool for Annotated Variants (ATAV). A minimum coverage of 10x was applied at the variant level. Prior to running the collapsing analysis, exons were excluded from the analysis if the case and control groups differed substantially in extent of coverage, specifically if one had on average ≥20% more of the exon real-estate adequately covered during sequencing than the other group. Following this "exon-pruning" step, any genes still showing a ≥3% difference in genic real-estate coverage between the cases and controls were removed (n=0). Only variants with a GATK value of pass or intermediate were included (variants predicted by VQSLOD as "fail" were excluded). To ensure that only high confidence variants were analyzed, variants that did not pass the additional QC metrics established above were excluded by providing a list of custom variants with ATAV's --exclude-variants command.

5.3 Results

5.3.1 Copy number variants in epilepsy families

Prior to selecting the multiplex epilepsy families for NGS sequencing, we screened for the presence of potentially pathogenic CNVs.

In Discovery Cohort A, a single proband from each of 88 families was screened for CNVs using a SNP array (see Methods). After quality control filtering, 880 CNVs were identified. Additionally, no previously known epilepsy genes[27] were included in any of these CNVs. One proband was found to harbor a CNV in the known pathogenic size range (>1MB). This CNV is a heterozygous deletion at 1q21.1, a CNV known to be associated with a wide variety of phenotypes, including epilepsy[32]. This CNV was not present in the other affected sibling in family and thus the pathogenicity of this CNV remains unclear in this family.

One proband from a different family harbored a rare heterozygous deletion at 4q28.3. An inherited homozygous deletion at this locus, not found in >2,000 controls, was reported in an autism patient[171]. Another rare deletion is reported in a schizophrenia patient, although it was inherited from an unaffected parent[172]. The heterozygous deletion in the proband from Discovery Cohort A is ~273kb and has slightly different boundaries than either of the previously reported 4q28.3 CNVs. This deletion includes a small protein coding gene, poly (A) binding protein, cytoplasmic 4-like (*PABPC4L*), of unknown function. Interestingly, heterozygous deletions in this region (non-identical boundaries) are also observed in 20 of 4,213 patients from the Epigen cohort, although deletions in this region are also found in controls and thus the enrichment in the sporadic epilepsy cohort is not significant (Fisher's exact test, two-tailed p=0.23). However, we cannot eliminate it as a possible susceptibility variant with

low penetrance in sporadic epilepsies and high penetrance within this family, especially because a TaqMan assay has revealed that this deletion segregates with epilepsy in this family (Figure 36).

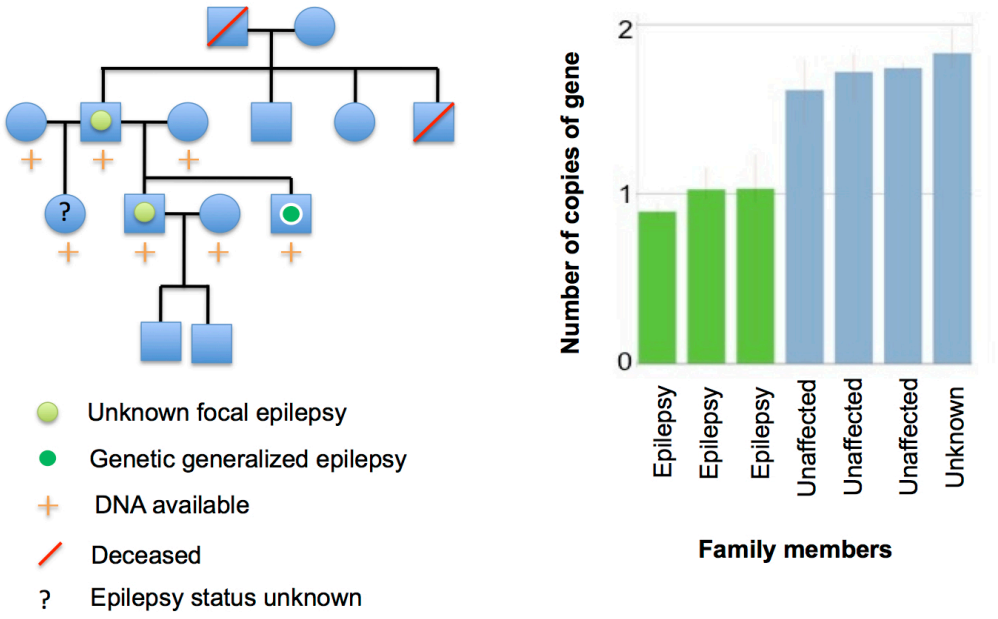


Figure 36. Cosegregation testing of 4q28.3 deletion in family 87001. Pedigree for family 87001 and the results of a TaqMan assay for the 4q28.3 heterozygous deletion. The 4q28.3 deletion segregates with epilepsy when testing all seven family members with DNA, including the relative of unknown status.

In Discovery Cohort B, we genotyped all 21 samples selected for exome-sequencing. However, sample bkv5912a2 failed genotyping and thus was excluded from subsequent CNV analyses. Since more than one affected relative in each of the ten families was screened, we also looked for CNVs present in both cases that were absent from the Database of Genomic Variants. We defined a "rare" CNV as one not present (<60% overlap) in the DGV database. The largest heterozygous deletion was ~780 Kb and

the largest homozygous deletion was ~98 Kb, both are considered common. A total of 73 rare heterozygous deletions were identified in the 20 Discovery Cohort B samples. The largest rare deletion was only 500kb and no rare homozygous deletions include any known genes. To identify CNVs that were shared by both family members, we defined sharing as CNVs with an overlap of >300bp. The majority of the shared rare CNVs did not intrupt any known genes and thus seem unlikely to be pathogenic. Another large number of these CNVs do not appear to actually be rare when checked against our other PennCNV called CHGV control samples (n=1,295). There are five remaining rare shared CNVs (Table 25).

Table 25. Rare shared CNVs in Discovery Cohort B sequenced samples.

Family	Sample	Gender	CNV	CN	Gene(s) within CNV	Overlap Sample	Gender	Overlaps with	CN	Size of Overlap (bp)
F	bkv6648f2	F	chrX:110851010-110851339	3	ALG13	bkv6622f1	F	chrX:110851010-110851339	3	330
J	bkv7065j1	F	chrX:110851010-110851339	3	ALG13	bkv7068j2	F	chrX:110851010-110851339	3	330
H	bkv19260h1	M	chr10:75121724-75131193	3	CTGLF7	bkv19263h2	M	chr10:75121724-75170290	3	9,470
J	bkv7065j1	F	chr21:40594402-40596684	1	DSCAM	bkv7068j2	F	chr21:40594402-40596684	1	2,283
G	bkv715g1	M	chr7:151592241-151630088	1	MLL3	bkv777g2	M	chr7:151592241-151630088	1	37,848

De novo mutations in *ALG13* have recently been associated with epileptic encephalopathies[42]; however the observed duplications only span an intronic region, making the pathogenicity of this CNV unclear. The heterozygous deletion spanning the "Down syndrome cell adhesion molecule" (*DSCAM*) gene is a potentially interesting

candidate gene; although, there was not enough evidence to put exome sequencing on hold for the two samples from this family. Interestingly, each of the case genomes from this family (bkvj) also contains a unique functional variant in the *DSCAM* gene (outside deletion) that is absent in controls. Thus, both family members may harbor a compound heterozygote in *DSCAM*. However, additional candidate variants are also found in this family making the impact of this experimentally unvalidated CNV unclear.

In total we screened 98 families for potentially pathogenic CNVs. This is the largest known study of the role of rare copy number variants in multiplex epilepsy families and reveals that none of the three established recurrent epilepsy CNVs can explain all instances of epilepsy in any of these families. Furthermore, no novel CNVs were recurrent within these families. While two families have candidate CNVs, it is likely that CNVs, at least those sensitive to these detection methods, do not play a major role in the etiology of these complex familial epilepsies.

5.3.2 Sequencing design and candidate variant identification

5.3.2.1 Sequencing design

To identify novel epilepsy susceptibility variants, whole-genome or exome sequencing was performed in each of 39 multiplex epilepsy families. Sequencing all family members or only affected family members was cost prohibitive. Given our primary hypothesis that a single variant will explain all instances of epilepsy in at least a subset of these multiplex families, we chose one predominant sequencing strategy. This

strategy was to sequence two affected and distantly related family members within a family to reduce the number of shared variants that are present as a result of identity by descent, while still identifying shared susceptibility variants. This novel sequencing design was used in ~60% (n=23) of the families (Figure 37). In two families we sequenced a third affected individual. In two other families we supplemented the sequencing of the original two whole-genome sequenced cases by exome sequencing of additional family members. Finally, in Discovery Cohort A, we identified 12 families in which all affected individuals suffered only from generalized seizures. Given that these genetic generalized epilepsy families might be under shared genetic control, we selected only a single affected individual to WGS from each of these "pure" GGE families. The majority of familial samples from the discovery cohorts were whole-genome sequenced with 13 families having at least two exome sequenced samples (Figure 37).

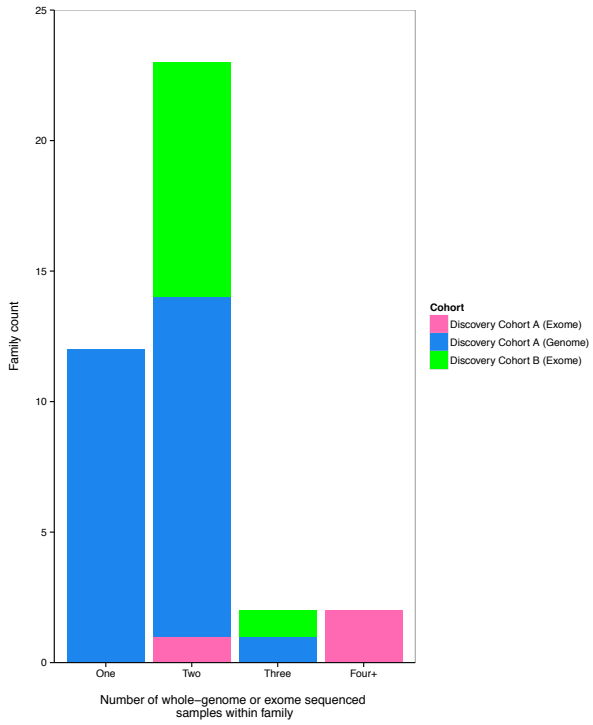


Figure 37. Sequencing strategies applied to 39 multiplex epilepsy families.

5.3.2.2 Candidate variant identification

For the familial epilepsy samples candidate variants were identified as those that were shared by all sequenced affected relatives that remained rare in control samples. A control MAF of 3% was applied to these samples due to the ~3% cumulative lifetime incidence of epilepsy, however variants found at lower frequencies in controls were prioritized. Despite having whole genome sequence data for some of our cases, we primarily focused on variants that were annotated as functional: nonsynonymous, truncating, variants in essential splice sites, or small insertion/deletion variants (indels)

located in the protein-coding regions. The number of candidate variants is effectively reduced with more distantly related relatives sharing fewer rare variants (Figure 38).

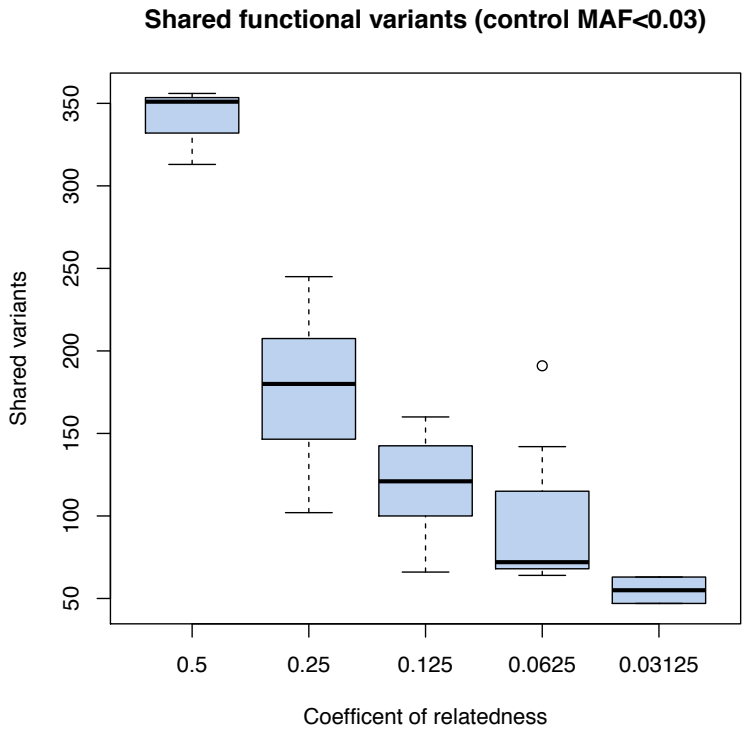


Figure 38. The number of rare shared functional variants decreases with increasingly distant relatives.

In addition to looking at variants shared within a family, it is possible to look across families to gain additional evidence for a given variant. When looking at candidate variants meeting the criteria explained above (and further restricted by rarity based on 387 controls for exon captured regions) in the 39 discovery families, we see ~5.6% of case identified variants showing up in multiple families (min 2, max 7 families).

If the control MAF cutoff is decreased to 1% then 3.14% of identified variants are seen in two or more families and when the control MAF cutoff is decreased to 0% then 0.89% of identified variants are seen in two or more families (min 2, max 3 families). Such variants were indirectly included in two ways: (i) if they were found in one of the initial 20 discovery families then they were genotyped in the large case-control cohort and (ii) such variants counted towards the aggregate gene rankings used in the resequencing experiment.

5.3.3 Association testing of candidate variants in a case-control cohort

For the familial epilepsy candidate variants, we further assessed variant frequency in a cohort of 949 sporadic epilepsy cases and 1,818 neurologically normal controls with the goal of identifying any variants significantly enriched in epilepsy cases. The number of control samples available at this time was very limited and databases such as the NHLBI ESP were not yet available; therefore assessing the control MAF in this large set of controls was very informative.

5.3.3.1 Selection of candidate variants for association testing

The selection of candidate variants from Discovery Cohorts A and B were very similar with a few small differences as explained below. These differences were mostly due to being selected at different times with differing resources and availability of real-estate on the custom iSelect genotyping chip.

In order to identify candidate variants in Discovery Cohort A, we looked for shared variants in the first ten sequenced families (otepia-otepij). We defined a candidate variant as: (i) having a control MAF ≤0.15% based on being present in no more than one of 331 exome sequenced unrelated controls (variants in non-exome captured regions were filtered based on being present in no more than one out of 68 whole-genome sequenced unrelated controls), (ii) functional SNV (intron-exon boundary, essential splice site, splice site, nonsynonymous, stop gain, stop loss), and (iii) shared by all sequenced members of a family. Two variants were removed after initial filtering because HapMap frequencies were >5% in the CEU population. Additionally, a variant in one of these families (otepii) was identified as "possibly shared" due to low (<10x) coverage in one of the two sequenced samples, this variant was in a biologically interesting gene (*MAG*), seen in two unrelated epilepsy cases, and in a region with a low linkage peak (LOD < 3.0) for this family. No other "possibly shared" variants were included. This resulted in the selection of 368 candidate variants that were selected for inclusion on the iSelect custom genotyping chip (286 variants passed the design phase).

In order to identify candidate variants in Discovery Cohort B, we looked for shared variants in the first nine sequenced families (sequence data was not available for bkvf at this time). We defined a candidate variant as: (i) having a control MAF ≤0.14% based on being present in no more than one of 362 whole-genome or exome sequenced unrelated controls, (ii) functional (essential splice site, splice site, nonsynonymous, stop

gain, stop loss, frameshift indels and coding-disrupting indels), (iii) in a gene with an HGNC recognized name (http://www.genenames.org) and (iv) shared by all sequenced members of a family. This resulted in the selection of 376 candidate variants that were selected for inclusion on the iSelect custom genotyping chip (323 variants passed the design phase).

5.3.3.2 Association testing results for familial variants assessed in a large case-control cohort

To formally test for enrichment of variants in epilepsy cases as compared to controls, we first adjusted for population stratification using the EIGENSTRAT program[169] and then used the resulting eigenvectors as covariates in the logistic regression (allelic model). Thus variant frequency differences were tested in a subset of genotyped samples, including 520 cases and 869 controls of European descent. The resulting logistic regression p-values were plotted using a quantile-quantile plot (Q-Q plot) and appear to have successfully been adjusted to account for population stratification (Figure 39).

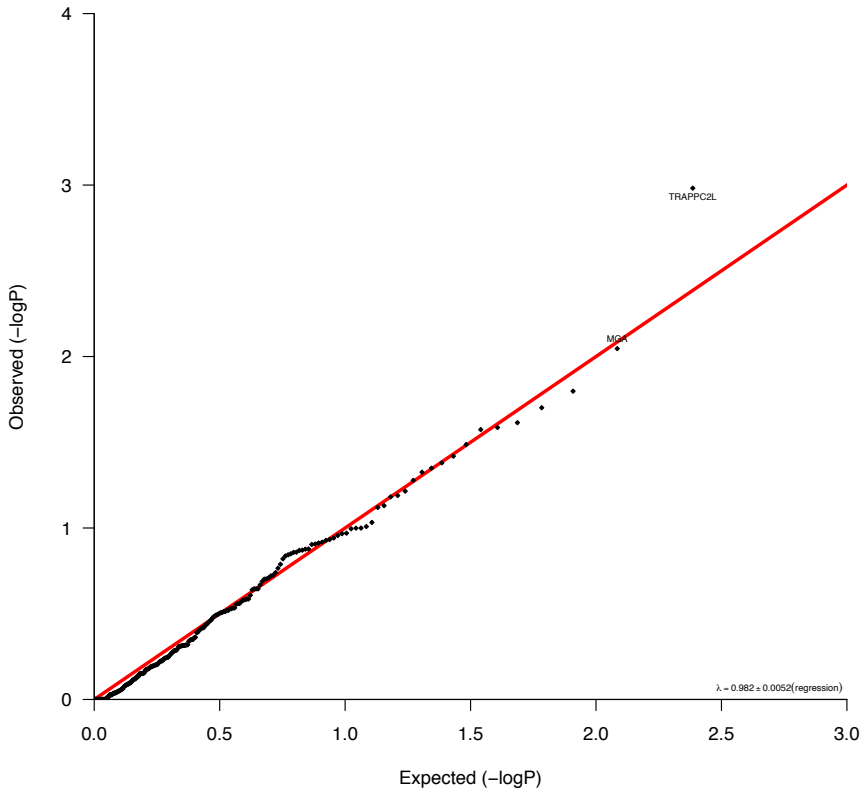


Figure 39. The quantile-quantile plot for familial epilepsy variants tested for enrichment in a population stratification corrected case-control cohort.

The variants with the lowest p-values are listed in Table 26; however none of the tested variants (n=609) were significantly enriched in epilepsy cases after Bonferonni correction for all tested variants.

We also looked for variants found exclusively in epilepsy cases and absent in controls after genotyping in the large case-control cohort. A total of 33 variants were successfully genotyped and found in the original discovery family and one additional epilepsy cases (n=949 sporadic cases) and absent in 1,818 controls (Table 27).

Table 26. Top 10 variants from a logistic regression using 520 unrelated epilepsy cases and 869 unrelated controls (all samples are of European ancestry).

No variants reached significance after Bonferroni correction for all tested variants. Variant IDs are based on hg18 coordinates (chr_coordinate_allele); no variants were present in dbSNP 129; the annotated function is based on the listed transcript ID (Ensembl 50_36I).

Variant ID	Discovery Family	Gene	Transcript	Annotated Function	Uncorrected p-value	Corrected p-value
16_87452613_G	bkvj	TRAPPC2L	ENST00000301021	Nonsynonymous	0.001043	NS
15_39829015_G	multiple families	MGA	ENST00000389936	Nonsynonymous	0.008992	NS
9_112209440_G	bkve	SVEP1	ENST00000297826	Nonsynonymous	0.015930	NS
15_81478556_G	otepii	BTBD1	ENST00000379403	Nonsynonymous	0.019890	NS
19_41809288_G	bkvj	ZNF382	ENST00000292928	Nonsynonymous	0.024310	NS
16_74121336_T	bkvi	CHST5	ENST00000336257	Nonsynonymous	0.025950	NS
1_22039047_G	otepic	HSPG2	ENST00000374695	Nonsynonymous	0.026680	NS
7_105041618_T	multiple families	ATXN7L1	ENST00000388807	Nonsynonymous	0.032620	NS
1_243917136_G	multiple families	KIF26B	ENST00000406578	Nonsynonymous	0.038130	NS
20_30562924_A	multiple families	C20orf112	ENST00000375673	Intron-exon boundary	0.041700	NS

Table 27. Familial variants absent in controls and present in more than one unrelated epilepsy sample.

The number of unrelated epilepsy carriers includes the positive control sample from the original discovery family. All variants are heterozygous. Variant IDs are based on hg18 coordinates (chr_coordinate_allele); the annotated functions are abbreviated as nonsynonymous (NS), intron-exon boundary (IEB), essential splice site (ESS), and stop gained (SG).

Cohort	Family ID	Gene	Variant ID	Function	PolyPhen-2	Total number of unrelated epilepsy case carriers
B	bkvj	KIF13A	6_17872374_A	NS	possibly	4
B	bkvi	CDON	11_125356153_A	ESS	n/a	3
A	otepii	SLC12A7	5_1128488_A	IEB	n/a	3
A	otepii	USP45	6_99991967_T	SG	n/a	3
B	bkvc	MDC1	6_30788277_C	NS	possibly	3
B	bkvi	LPL	8_19856013_A	NS	probably	3
B	bkvi	MYCBP2	13_76597556_T	NS	probably	3
A	otepii	CLK3	15_72699420_G	NS	possibly	2
A	otepib	PLA2R1	2_160597760_A	NS	possibly	2
A	otepic	APAF1	12_97580435_T	NS	probably	2
A	otepib	PFKM	12_46813506_G	NS	probably	2
A	otepic	TYR	11_88564113_A	NS	probably	2
B	bkvb	FYCO1	3_45983588_C	NS	benign	2
B	bkvd	PCCB	3_137530312_G	NS	benign	2
B	bkvc	TTC16	9_129519427_T	NS	benign	2
B	bkvj	USP37	2_219122867_A	NS	benign	2
B	bkvd	DNAJC13	3_133700702_A	NS	benign	2
B	bkvj	TMEM128	4_4298855_C	NS	benign	2
A	otepic	MGAT4A	2_98626930_A	NS	benign	2
A	otepib	SPATS2	12_48204939_C	NS	benign	2
B	bkvb	AL592464.24	1_2969429_T	NS	unknown	2
B	bkvd	COMP	19_18756131_T	NS	possibly	2
B	bkvc	DBC1	9_120969614_T	NS	possibly	2
B	bkve	EXTL1	1_26221774_T	NS	possibly	2
B	bkvc	ITGA4	2_182095214_A	NS	possibly	2
B	bkvj	ABCB5	7_20729333_G	NS	probably	2
B	bkvd	CHRNA9	4_40045712_A	NS	probably	2
B	bkvc	CPXM2	10_125612160_A	NS	probably	2
B	bkvj	FAM113B	12_45916306_T	NS	probably	2
B	bkvd	NOL8	9_94117023_C	NS	probably	2
B	bkvd	PPP2R3A	3_137307776_T	NS	probably	2
B	bkvi	RHOT2	16_662959_T	NS	probably	2
A	otepib	COL5A1	9_136834562_T	NS	unknown	2

The variant found in the most cases is found in three unrelated epilepsy cases (plus the original discovery family) and causes a nonsynonymous coding change in *KIF13A*. This variant shows cosegregation within the discovery family and was included in the resequenced genes where we observe a rare functional variant in an additional family from Replication Cohort D. Despite these 33 variants not reaching significance in this study, these variants and the genes harboring these variants, are strong epilepsy susceptibility candidates.

5.3.4 Cosegregation of candidate variants

We tested cosegregation of candidate variants identified by whole-genome or exome sequencing in 19 of the families from the discovery cohorts. For Discovery Cohort A, we comprehensively evaluated all candidates (rare, shared, and functional variants included on the iSelect custom genotyping chip). These candidates were evaluated by including additional family members on the iSelect genotyping chip. Therefore, 216 candidate variants were genotyped for cosegregation in the original nine Discovery Cohort A families. The number of variants genotyped in each family was directly correlated with the relatedness of the whole-genome sequenced samples, consequently, so was the number of candidate variants identified (Figure 38); this ranged from 5 to 41 variants per family. For nine of the ten families from Discovery Cohort B, we first comprehensively evaluated all candidates for their frequency in a larger control cohort (rare, shared, and functional variants included on the iSelect custom genotyping chip)

and then selected a subset of the strongest candidates for cosegregation testing by Sanger sequencing. The 10th family did not have NGS data prior to the iSelect experiment. The subset of strongest candidates was selected by prioritizing the rarest variants (based on original ~370 CHGV sequenced controls, ~1,800 CHGV iSelect genotyped controls [for nine of the ten families], and ~2,500 available NHBLI ESP controls), nonsynonymous variants with a PolyPhen-2 prediction of benign were excluded, genes were prioritized if shared rare functional variants were also found in multiple epilepsy families, and finally, the known biological functions of the gene were considered. This resulted in successful sequencing of 39 variants across these 10 families (range one to six per family). In total, 255 variants were successfully genotyped within the relevant discovery family (n=19).

We then analyzed the resulting genotype of each variant, for the available samples, within the relevant discovery family. Given the complex inheritance patterns observed in these pedigrees, all unaffected individuals were allowed to be variant carriers – indicating decreased penetrance of the variant. To assess if a variant segregates with disease in these pedigrees we assumed two models. The primary model used was "all or none" cosegregation in which all individuals with epilepsy were required to carry the variant. The only exceptions to this were individuals with symptomatic (structural metabolic) epilepsy or individuals with an uncertain epilepsy diagnosis. For a small number of these families we also considered a "best fit" cosegregation model where one

of several scenarios was observed: (i) a variant segregated with epilepsy in only one branch of the larger pedigree, (ii) a variant segregated in a phenotype-specific manner, or (iii) a variant showed a constellation within a pedigree with a complex combination of phenotypes that suggested the variant may explain some but not all instances of epilepsy.

In total, 81 variants showed cosegregation within their discovery family. Approximately, 65% of these variants cosegregated according to the all or none model. On average, 4.5 variants cosegregated per family and for variants that remained completely absent in controls, this average was 3.1. The subset of variants that cosegregate by the all or none model and remain completely absent in controls (n=28) are a very strong set of candidates and there are 11 families with such variants; with two families having a single variant in this category and three families having four variants in this category. Details for all cosegregating variants are shown in Table 29.

By the all or none model, there are only two families with a single cosegregating variant. Neither of these two variants is convincingly causal. The variant in the first family (otepia) is in *ZNF630* residing on the X-chromosome and, since both males and females are equally affected, suggests X-linked dominant inheritance. However, since the father is affected it is inferred that he is the carrier, which would result in none of his sons having the disease; instead, five of his six sons all have epilepsy. Therefore, while this variant technically cosegregates with disease, the mode of inheritance is unlikely. In

the second family (otepid), the cosegregating variant is a nonysnonymous variant found in a genomic region predicted to encode (AC097370.10) an uncharacterized protein. Additionally, without application of the "best fit" model for cosegregation, six (~32%) of these families would have no segregating variants.

Table 28. Variants showing cosegregation in one of the 19 tested families from Discovery Cohort A or B.

For each family, the total numbers of tested variants are listed along with the number of variants showing cosegregation by the "all or none" or "best fit" model. Variant IDs are based on hg18 coordinates (chr_coordinate_allele). The annotated functions are abbreviated as nonsynonymous (NS), intron-exon boundary (IEB), stop gained (SG), and frameshift (FS).

Family	Variant ID	Gene	Function	Control MAF 0%	Cosegregation results	Tested variant count	All or none	Best fit
otepia	1_94245817_A	ABCA4	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	36	1	10
otepia	14_91695294_C	CPSF2	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	36	1	10
otepia	1_34270836_A	CSMD2	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	36	1	10
otepia	18_13875316_C	MC2R	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	36	1	10
otepia	X_47721726_G	ZNF630	NS	N	Cosegregates	36	1	10
otepia	3_150097018_C	CPA3	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	36	1	10
otepia	21_36502956_A	DOPEY2	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	36	1	10
otepia	22_22245697_G	IGLL1	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	36	1	10
otepia	20_5853630_A	CHGB	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	36	1	10
otepia	22_35011140_C	MYH9	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	36	1	10
otepia	3_150676360_A	TM4SF4	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	36	1	10
otepib	6_34323430_T	C6orf1	IEB	N	Cosegregates	41	10	0
otepib	12_48204939_C	SPATS2	NS	Y	Cosegregates	41	10	0
otepib	12_48246781_C	MCRS1	NS	Y	Cosegregates	41	10	0
otepib	15_39924063_A	PLA2G4B	NS	N	Cosegregates	41	10	0
otepib	22_39935722_C	L3MBTL2	NS	N	Cosegregates	41	10	0
otepib	22_24634320_A	MYO18B	NS	Y	Cosegregates	41	10	0
otepib	15_39921031_A	PLA2G4B	NS	N	Cosegregates	41	10	0
otepib	12_51332094_A	KRT2	NS	N	Cosegregates	41	10	0
otepib	15_38887772_T	ZFYVE19	NS	N	Cosegregates	41	10	0
otepib	15_39930579_A	SPTBN5	NS	N	Cosegregates	41	10	0
otepic	9_124194423_A	PTGS1	NS	N	Cosegregates	30	4	0
otepic	12_97580435_T	APAF1	NS	Y	Cosegregates	30	4	0
otepic	16_4347229_A	CORO7	NS	Y	Cosegregates	30	4	0

Family	Variant ID	Gene	Function	Control MAF 0%	Cosegregation results	Tested variant count	All or none	Best fit
otepic	16_1852031_C	C16orf73	NS	Y	Cosegregates	30	4	0
otepid	5_161061190_C	GABRA6	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	14	1	2
otepid	13_22805481_G	SACS	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	14	1	2
otepid	17_3004475_T	AC097370.10	NS	Y	Cosegregates	14	1	2
otepif	9_94025441_A	IARS	IEB	Y	Cosegregates	20	5	0
otepif	9_138386293_A	CARD9	NS	Y	Cosegregates	20	5	0
otepif	9_138848023_G	C9orf86	NS	N	Cosegregates	20	5	0
otepif	9_124699425_C	RC3H2	NS	Y	Cosegregates	20	5	0
otepif	3_15474737_A	COLQ	NS	Y	Cosegregates	20	5	0
otepig	17_1363908_T	SKIP	IEB	N	Cosegregates	10	5	0
otepig	2_64535982_G	AC008074.1	NS	Y	Cosegregates	10	5	0
otepig	16_49731730_A	SALL1	NS	N	Cosegregates	10	5	0
otepig	16_57151661_A	CNOT1	NS	Y	Cosegregates	10	5	0
otepig	16_47969973_A	C16orf78	NS	N	Cosegregates	10	5	0
otepih	11_35452760_C	AC090625.5	NS	N	Cosegregates	19	8	0
otepih	2_166243861_G	FAM130A2	NS	N	Cosegregates	19	8	0
otepih	11_77089788_A	RSF1	NS	N	Cosegregates	19	8	0
otepih	11_62194056_C	C11orf48	NS	Y	Cosegregates	19	8	0
otepih	16_2033640_A	NTHL1	NS	Y	Cosegregates	19	8	0
otepih	10_34665168_G	PARD3	NS	Y	Cosegregates	19	8	0
otepih	16_31180597_A	ITGAM	NS	N	Cosegregates	19	8	0
otepih	2_15473237_T	NAG	NS	Y	Cosegregates	19	8	0
otepii	3_185491349_A	ECE2	IEB	N	Cosegregates	41	11	0
otepii	15_50449773_T	MYO5A	NS	N	Cosegregates	41	11	0
otepii	1_203763575_A	PCTK3	NS	Y	Cosegregates	41	11	0
otepii	3_197765553_C	WDR53	NS	Y	Cosegregates	41	11	0
otepii	19_41265856_T	WDR62	NS	N	Cosegregates	41	11	0
otepii	7_151580628_G	MLL3	NS	Y	Cosegregates	41	11	0
otepii	6_111795640_A	REV3L	NS	N	Cosegregates	41	11	0

Family	Variant ID	Gene	Function	Control MAF 0%	Cosegregation results	Tested variant count	All or none	Best fit
otepii	1_198886237_T	DDX59	NS	N	Cosegregates	41	11	0
otepii	13_109915924_T	COL4A2	NS	N	Cosegregates	41	11	0
otepii	4_9393756_A	DRD5	SG	N	Cosegregates	41	11	0
otepii	6_99991967_T	USP45	SG	Y	Cosegregates	41	11	0
bkva	4_122180609_A	C4orf31	NS	N	Cosegregates with all focal epilepsy cases	1	0	1
bkvb	7_125960765_A	GRM8	NS	Y	Cosegregates	5	2	0
bkvb	2_220207542_T	SLC4A3	NS	Y	Cosegregates	5	2	0
bkvc	9_130419850_T	SPTAN1	NS	Y	Cosegregates	5	2	0
bkvc	1_64820458_A	CACHD1	NS	N	Cosegregates	5	2	0
bkvd	18_19369497_INS_T	NPC1	FS	Y	Cosegregates in one branch of the family	6	0	4
bkvd	2_202297425_C	ALS2	NS	Y	Cosegregates in one branch of the family	6	0	4
bkvd	4_40045712_A	CHRNA9	NS	N	Cosegregates in one branch of the family	6	0	4
bkvd	3_38810465_A	SCN10A	NS	N	Cosegregates in one branch of the family	6	0	4
bkve	1_26221774_T	EXTL1	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	5	0	3
bkve	9_119516296_G	TLR4	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	5	0	3
bkve	8_81896206_C	ZNF704	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	5	0	3
bkvf	12_108410848_A	UBE3B	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	3	0	2
bkvf	3_49314832_A	USP4	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	3	0	2
bkvg	17_75541121_T	TBC1D16	NS	Y	Cosegregates in one branch of the family	2	0	1
bkvh	1_113054364_T	PPM1J	NS	N	Evidence for cosegregation but complex phenotypes and inheritance	2	0	2
bkvh	10_129132450_T	DOCK1	NS	N	Evidence for cosegregation but complex phenotypes and inheritance	2	0	2
bkvi	7_34854754_T	NPSR1	NS	Y	Cosegregates	5	2	2
bkvi	20_36790521_C	SLC32A1	NS	Y	Cosegregates	5	2	2
bkvi	13_76597556_T	MYCBP2	NS	N	Cosegregates with generalized epilepsy cases	5	2	2
bkvi	1_43686205_A	KIAA0467	NS	Y	Cosegregates with generalized epilepsy cases	5	2	2
bkvj	6_17872374_A	KIF13A	NS	Y	Cosegregates	5	2	1
bkvj	2_202330576_G	ALS2	NS	Y	Evidence for cosegregation but complex phenotypes and inheritance	5	2	1
bkvj	5_177629945_A	COL23A1	NS	Y	Cosegregates	5	2	1

Cosegregation testing in these 19 epilepsy families emphasizes that evidence of a cosegregating variant within a single family is not sufficient to prove causality. While cosegregation testing can be formalized by using linkage programs such as Merlin[173] to calculate a LOD score and corresponding p-value, this will not withstand correction for the number of tests conducted. Furthermore, all but two of these families had more than one variant showing cosegregation indicating that independent evidence is needed to decipher which variant(s) is causal.

5.3.5 Candidate gene resequencing in a large familial cohort

A total of 267 families comprise Replication Cohort D. We used this cohort to resequence the 68 "top" candidate genes associated with familial epilepsies. Again, we primarily employed the sequencing strategy of sequencing multiple individuals per family resulting in 200 families with two sequenced affected relatives, three families with three sequenced affected family members and 64 families with a single proband sequenced. In total, 473 familial epilepsy cases were successfully sequenced for this experiment.

5.3.5.1 Selection of genes for resequencing

The Replication Cohort D families were not systematically screened for mutations in the phenotypically relevant genes. Therefore, we first included seven known epilepsy genes that were most relevant to the familial phenotypes of Replication Cohort D (Table 29).

Table 29. The seven known epilepsy genes included in resequencing experiment.

GENE	Associated Phenotype(s) [Phenotype MIM number]
LGI1	Epilepsy, familial temporal lobe, 1 [600512]
GABRD	Epilepsy, idiopathic generalized, 10 [613060]
GABRG2	Epilepsy, generalized, with febrile seizures plus, type 3 [611277]
SCN1A	Dravet syndrome [607208]; Epilepsy, generalized, with febrile seizures plus, type 2 [604403]; Migraine, familial hemiplegic, 3 [609634]
SCN2A	Epileptic encephalopathy, early infantile, 11 [613721]; Seizures, benign familial infantile, 3 [607745]
SCN1B	Epilepsy, generalized, with febrile seizures plus, type 1 [604233]
SCN9A	Epilepsy, generalized, with febrile seizures plus, type 7 [613863]

Next, we identified candidate genes as those harboring rare functional variants in Discovery Cohorts A and B (39 families). Association testing at the gene-level (collapsing analysis) was not sufficiently powered with only 39 cases. In order to obtain a list of candidate genes, we looked for genes with qualifying variants in multiple epilepsy families. A qualifying variant was defined as: (i) having a control MAF <1% based on 387 whole-genome or exome sequenced unrelated controls, (ii) functional, and (iii) shared by all sequenced members of a family (or present the sole sequenced proband). Across all 39 families, 4,486 genes harbor a qualifying variant in ≥1 family. Only 1,414 genes harbor a qualifying variant in ≥2 families and 450 genes harbor a qualifying variant in ≥3 families. Two genes, *MUC4* and *TTN*, harbored qualifying variants in 16 and 14 families, respectively. Despite these genes showing up in the largest number of families, they are also some of the largest genes in the genome and are

found in the 98th and 99th percentile of intolerant genes [124] indicating that these genes are very tolerant to functional variation. Given that genes causing Mendelian disorders are more intolerant to functional variation [124], one approach to selecting the top candidate epilepsy genes from this subset would be to simply apply the intolerance scoring system and take the genes most intolerant to functional variation. However, the epilepsy phenotypes in these families are arguably less severe than many Mendelian diseases and definitely less severe than the epileptic encephalopathies, making it difficult to know where to set a strict RVIS threshold. In fact, some of the genes causing neurodevelopmental disorders are tolerant to functional variation (Figure 25) and two of the well-established epilepsy genes selected for inclusion in this resequencing experiment (*SCN1B* and *SCN9A*) are considered tolerant to functional variation (in the 83rd and 73rd percentile, respectively).

Therefore, we used a combined approach for candidate gene selection by incorporating the tolerability score for a gene (RVIS) [124], biological evidence for the role of a gene in epilepsy (PVEG score, see methods), and finally, the number of families with qualifying variants in a gene. This decision tree (Figure 40) led to the selection of 68 genes for resequencing. Three or more families harbor a qualifying variant in 31 of the selected 68 genes.

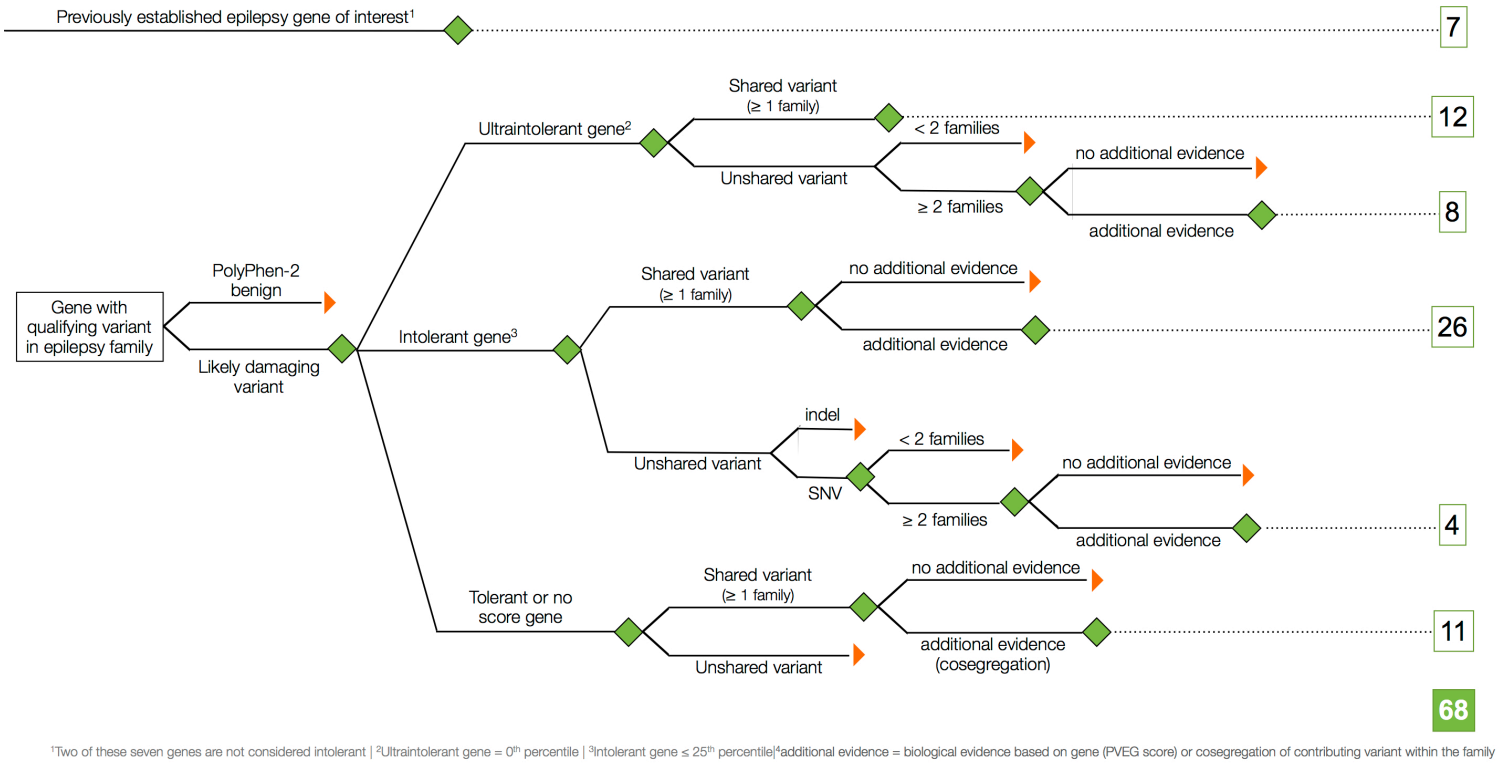


Figure 40. Decision tree for selection of 68 candidate epilepsy genes

Additional evidence is based on a PVEG score ≥ 1 and/or evidence of cosegregation of the qualifying variant in the discovery family. [1]Two of these seven genes are considered tolerant (RVIS >25th percentile). [2]Ultraintolerant genes have an RVIS in the 0th percentile. [3]Intolerant genes have an RVIS ≤ 25th percentile.

5.3.5.2 Identification of explained epilepsy families

By focusing on the seven previously established epilepsy genes of interest, we identified "solved" families using a strict set of filtering criteria. Variants in these genes were filtered to only include those which met the following criteria: (i) functional, (ii) control MAF <1% (iii) shared by all sequenced members of a family (or in a family with a single sequenced proband), and (iv) the exact mutation was previously reported in the literature as disease-causing. Finally, the familial phenotype of the Replication Cohort D samples had to match that of the reported familial phenotype. This is a very conservative approach to identifying clearly pathogenic variants in these families. Using this conservative method we explain 6% (n=17) of these epilepsy families – with six of the seven genes explaining one or more family: *SCN1A* (6), *SCN1B* (5), *SCN9A* (2), *GABRD* (2), *GABRG2* (1), and *LGI1* (1).

There are other interesting candidate variants in these established epilepsy genes, which can likely explain a slightly higher percentage of families. For example, seven families have previously unreported variants meeting the following criteria: (i) frameshift, stop gain, or PolyPhen-2 probably damaging nonsynonymous, (ii) control MAF = 0% (iii) shared by all sequenced members of a family (or in a family with a single sequenced proband). Interestingly, two of these families have Familial Temporal Lobe Epilepsy (FTLE) where the qualifying variant is in *SCN1B* or *SCNA2* – mutations in these genes have not yet been associated with FTLE. Additionally, a number of

candidate variants were identified in one sequenced relative and not the other; reflecting two main possibilities: these variants were sometimes poorly covered at the variant site in only one of the two samples and thus might actually show full cosegregation in these families or the variant may show cosegregation with an isolated branch of this family, explaining some but not all instances of epilepsy. We are in the process of following up many of these candidate variants by cosegregation testing. Importantly, this work will likely clarify the phenotypic spectrum associated with mutations in the established epilepsy genes.

5.3.5.3 Gene-based collapsing for unexplained epilepsy families

Rather than testing variants individually, another approach is to "collapse" a set of variants across a genomic region and then test their collective frequency differences between cases and controls [174]. Here we collapse rare and functional variants across each of the 68 genes selected for resequencing. We do not have genome-wide genotyping data for Replication Cohort D samples and thus we cannot formally correct for population stratification; however, we did limit the tested samples to those of self-declared European ancestry (~80% of Replication Cohort D). Additionally, we removed the 17 explained epilepsy families. Including only a single relative from each of the remaining families resulted in 237 case samples and 1,565 European controls. Restricting to variants that are functional and completely absent in controls (MAF 0%) we identify two genes that reach significance after correcting for all 68 tested genes ($\propto=7.35\text{x}10^{-4}$

(0.05/68)) (Figure 41). *SCN1A* was the top gene with a p-value of 5.85×10^{-5} and *MYCBP2* was also significant with a p-value of 1.99×10^{-4}.

The significance of *SCN1A* in the gene-based collapsing analysis suggests that some additional families are likely explained by *SCN1A* mutations, but these families were not excluded based on the initial results from the analysis of variants in the established epilepsy genes. The *SCN1A* signal is driven by qualifying variants in ~5% of cases (12/225) and ~1% of controls (15/1,547).

The significance of *MYCBP2* in the gene-based collapsing analysis may reflect the discovery of a novel epilepsy susceptibility gene. The *MYCBP2* signal is driven by qualifying variants in ~6% of cases (15/222) and ~2% of controls (28/1,534) (Figure 42).

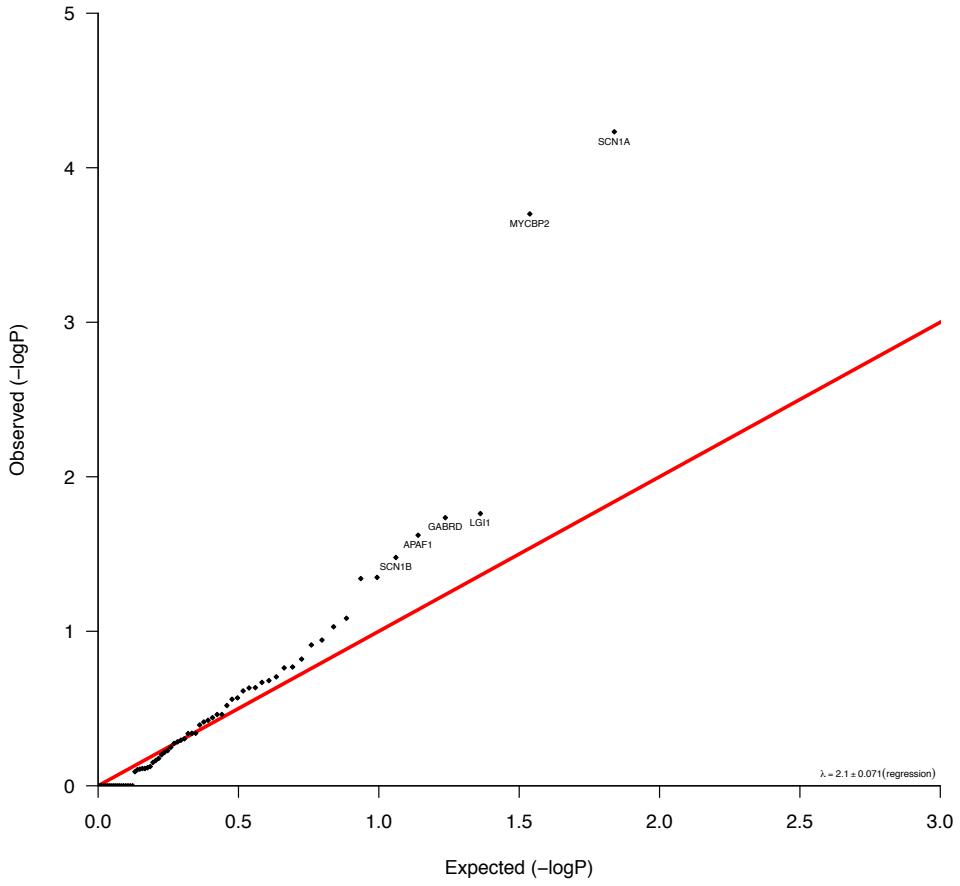


Figure 41. The quantile-quantile plot for the 68 genes resequenced in 237 European familial epilepsy probands and 1565 European controls.

MYCBP2 is highly expressed in the brain[175]; it encodes the MYC Binding Protein 2 and is predicted to be an E3 ubiquitin-protein ligase. The *MYCBP2* gene physically binds the neuron-specific electroneutral potassium and chloride cotransporter, *SLC12A5*[176]. Furthermore, *MYCBP2*'s drosophila ortholog has established the role of this protein not only in synaptic growth[177], but also its interaction with the TSC1-TSC2 complex associated with Tuberous Sclerosis (side effects

of which include seizures, ID, and autism). MYCBP2 colocalizes with TSC1 and TSC2 along the neurites and in the growth cones (http://www.uniprot.org/uniprot/O75592). Mice lacking the *MYCBP2* ortholog show neonatal lethality and die at birth (http://www.informatics.jax.org/allele/allgenoviews/MGI:3760640). These homozygous animals have gross morphological defects in the brain; including a thin corpus callosum, dilated lateral ventricles, hippocampal formations are reduced in size and dysmorphic. There is no current CCDS transcript for *MYCBP2* and thus this gene was not given an intolerance score[124]. However, I manually calculated the sum of all variant sites in the gene (X=317) and the sum of all common (MAF > 0.1%) variants in the gene (Y=8). The gene with the closest values is *INTS1* (X = 315, Y =7, RVIS percentile of 0.07), indicating that *MYCBP2* is also an extremely intolerant gene.

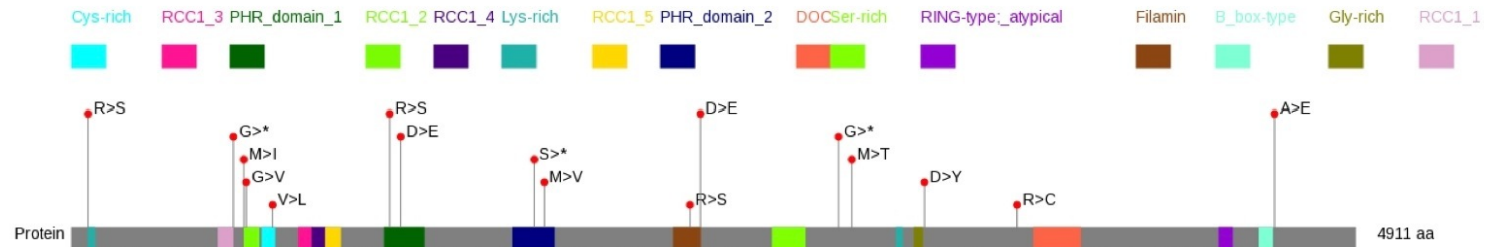
Qualifying variants from the gene-based collapsing analysis were plotted along UniProt domains of the MYCBP2 protein (Figure 42). There is no specific domain with an enrichment of case variants; however four of the 16 case variants fall within the two PHR domains, which are critical for proper localization and function[178].

MYCBP2 was selected for resequencing because it had several strong lines of evidence; including: cosegregation of the identified *MYCBP2* variant in a family from Discovery Cohort B (bkvi), two pure GGE cases from Discovery Cohort A (otepi10554x1 & otepi8884v1) harbored rare functional variants in *MYCBP2*, and a *de novo* mutation in *MYCBP2* was reported in one of the autism spectrum disorder (ASD) studies[58]. It has

also been suggested as an important node in ASD associated molecular pathways[58,179]. Additionally, the cosegregating variant from bkvi (13_76597556_T) was included in the iSelect experiment. This variant remained absent in 1,818 controls and this variant was also found in two unrelated nonfamilial GGE cases (Replication Cohort C) and two schizophrenia cases (combined epi+schiz p-value = 0.03, uncorrected).

In order to assess the evidence of pathogenicity, we are currently testing the identified *MYCBP2* variants in the relevant families.

15 qualifying case variants (6.3%)



28 qualifying control variants (1.8%)

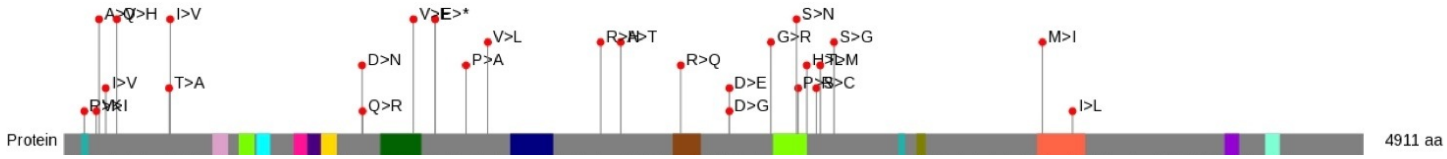


Figure 42. Location of qualifying variants (gene-based collapsing) in cases and controls across the MYCBP2 protein domains (UniProt).

5.3.5.4 Additional evidence for epilepsy susceptibility genes

Admittedly, one disadvantage to the case-control collapsing methods is that we only use a single familial case from each family and we are not considering the associated family structure until cosegregation testing of variants in the implicated genes. There are likely additional genes of interest included amongst the 68 resequenced genes. As a basic screen for such genes, the number of families harboring rare and functional variants can be used to rank the genes. First, we again removed the 17 explained epilepsy families. Second, we required the considered variants to be shared by all sequenced members of a family (or in a family with a single sequenced proband). Next, the qualifying variants were limited to those with a control MAF of zero and nonsynonymous variants that were predicted to be benign by PolyPhen-2 were excluded. Then only the most damaging variant per family per gene (nonsense, splice, nonsynonymous PolyPhen-2 predicted probably damaging, nonsynonymous PolyPhen-2 predicted possibly damaging) was kept. This resulted in 27 genes with a qualifying variant in ≥2 families (Table 30). Additional evidence will be needed to clarify the role of these genes in epilepsy risk.

Table 30. Genes with exceptionally rare and functional variants in two or more epilepsy families.

Genes are listed in descending order by the total number of families harboring a shared variant. RVIS percentiles [124] are shown in the last column with the 2% most intolerant genes highlighted in red text.

	GEFS+		Focal				
Gene	Families with shared variants	Families with single sample sequenced (unshared variant)	Families with shared variants	Families with single sample sequenced (unshared variant)	Total families with shared variants	TOTAL	RVIS %ile
'RYR1'	2	2	1	0	3	5	0.0059
'DOPEY2'	2	3	0	0	2	5	0.2772
'CACNA1H'	2	3	0	0	2	5	1.6277
'AHNAK'	2	2	0	0	2	4	0.6134
'SPEN'	2	2	0	0	2	4	1.2090
'CAD'	0	0	2	0	2	2	0.2300
'SCN1B'	1	0	1	0	2	2	83.2508
'MYO18B'	2	0	0	0	2	2	93.0231
'SACS'	0	6	1	0	1	7	2.0996
'VPS13A'	0	5	1	0	1	6	2.0347
'MYCBP2'	0	5	1	0	1	6	N/A
'COL7A1'	1	2	0	0	1	3	0.1297
'HMCN1'	1	2	0	0	1	3	0.2477
'ANK3'	1	2	0	0	1	3	0.3362
'SCN1A'	1	2	0	0	1	3	4.0281
'SCN10A'	1	2	0	0	1	3	4.7358
'TLN1'	0	1	1	0	1	2	0.3657
'SCN2A'	0	1	1	0	1	2	1.7693
'GRIN2A'	0	1	1	0	1	2	3.8865
'ITGAM'	1	1	0	0	1	2	18.2826
'APAF1'	0	1	1	0	1	2	47.2045
'LRP1'	0	0	1	0	1	1	0.0177
'SPTAN1'	1	0	0	0	1	1	0.3126
'HTT'	0	0	1	0	1	1	1.6926
'SLC4A3'	0	0	1	0	1	1	5.9153
'LGI1'	0	0	1	0	1	1	14.4020
'TREX1'	0	0	1	0	1	1	15.7643
'COLQ'	1	0	0	0	1	1	20.7006
'HERC2'	0	2	0	0	0	2	0.0472
'BSN'	0	2	0	0	0	2	0.1946

	GEFS+		Focal				
Gene	Families with shared variants	Families with single sample sequenced (unshared variant)	Families with shared variants	Families with single sample sequenced (unshared variant)	Total families with shared variants	TOTAL	RVIS %ile
'MCM3AP'	0	2	0	0	0	2	0.8669
'RELN'	0	2	0	0	0	2	1.4626
'ALS2'	0	2	0	0	0	2	3.8865
'DOCK1'	0	2	0	0	0	2	N/A
'TLN2'	0	1	0	0	0	1	0.3539
'CELSR1'	0	1	0	0	0	1	0.6310
'CNTNAP2'	0	1	0	0	0	1	1.2975
'UBE3B'	0	1	0	0	0	1	1.3152
'CSMD2'	0	1	0	0	0	1	1.9226
'CACHD1'	0	1	0	0	0	1	2.4180
'MYO5A'	0	1	0	0	0	1	4.1460
'GABRG2'	0	1	0	0	0	1	25.1474
'KIF13A'	0	1	0	0	0	1	46.4968
'NPC1'	0	1	0	0	0	1	50.0177
'SCN9A'	0	1	0	0	0	1	73.6259

5.4 Conclusions

We performed whole-genome or exome sequencing in 39 multiplex epilepsy families. We sought to uncover novel epilepsy susceptibility factors using both variant-based and gene-based approaches in larger follow up cohorts. However, we did not clearly implicate variants or genes associated with epilepsy, consistent with previous observations that the epilepsies (including Mendelian families and epileptic encephalopathies) have extremely high locus and allelic heterogeneity. Studying a small number of large multiplex families did not facilitate the identification of inherited risk factors for epilepsy; suggesting that very large sample sizes are needed. Family based designs can still be powerful in these large cohorts and additional approaches such as transmission disequilibrium testing will be feasible.

It is conceivable that these large multiplex families are enriched for families where independent variants segregate in different branches of the family and therefore, our selection of distantly related relatives to enrich for shared risk variants may have resulted in the opposite effect. The phenotypic heterogeneity within many of these families may reflect more than one genetic risk factor segregating throughout the family (although other models are also possible, such as variable seizure thresholds influenced by the environment); again possibly being missed by our approach of sequencing distantly related relatives which generally did not exclude discordant seizure phenotype pairings. If some proportion of these families do follow a monogenic model of inheritance, then sequencing pairs of distantly related relatives is an effective method for reducing the number of candidate variants and provides a more manageable number of variants for follow up. However, this work also emphasizes that caution should be used when attempting to prove causality of a variant given cosegregation of a damaging variant within a single pedigree.

This work has identified a set of candidates, among which, some will likely prove to be real risk factors for epilepsy. However, it will take more work to prove these as real risk factors for epilepsy. The most promising of these is the *MYCBP2* gene emerging from the large familial resequencing study. We are currently in the process of testing the contributing variants for cosegregation within the appropriate families.

Additional work with these genomes and increasing sample sizes will likely lead to the identification of new genetic variants involved in epilepsy, which will offer important insights into epileptogenesis and lead to the development of novel therapies. Finally, once causal variants are identified we can add pieces to the complex puzzle of the genotype-phenotype interactions of epilepsy.

6. Conclusions and future directions[1,2]

There is no question that next-generation sequencing has revolutionized human genetics research. Over the past six years, chemists, biologists, and engineers have continually improved the next-generation sequencing technology itself, making it cheaper and faster while maintaining or even improving quality (for example, longer read lengths). Furthermore, bioinformaticians have developed algorithms that can reliably identify single nucleotide variants and, somewhat less reliably, detect small insertions and deletions. Algorithms for detection of large deletions and duplications from whole-genome sequence data are considered well validated, while identification methods based on exome sequence are still in their infancy. We still struggle with detection of inversions and more complex rearrangements, and we need to address the "big data" problem in biology by developing superior computing power and more efficient ways to move the data around.

Human geneticists have had great success in the application of NGS to Mendelian diseases: from the first success in September of 2009 [181] through January 2011, novel genetic causes were identified and *published* for at least 30 disorders[182,183]. Not only is this speed impressive, but also that these discoveries included very small sample sizes and sporadic disorders, which are refractory to traditional linkage analysis.

---

[1] Portions of this chapter have been modified from a published work[180] .
[2] Portions of this chapter are being submitted for publication in a textbook for medical students entitled "Practical Epilepsy" to be published by Demos Medical Publishing.

I discovered the cause of ASNS deficiency in August of 2010, but this work was not published until October 2013. The original discovery was based only on Families A and B who carried the same homozygous mutation, meaning that there was a very small possibility that this was a rare variant in the Iranian Jewish population that was not causal. Thus, we embarked on functional work and characterization of a mouse model to gain evidence for pathogenicity. After presenting our work at a meeting (publicly available abstract) we were approached by a collaborator with two additional families with overlapping phenotypes and mutations in *ASNS*. It is likely that there are many stories like this that don't end in two united research labs and instead, many researchers are individually sitting on causal variants that can't be published. In the case of the undiagnosed clinical sequencing study (Chapter 3), we plan to publish the phenotypes and genetic findings for the "likely genetic diagnosis" cases in the hopes that additional patients can be recognized with overlapping phenotypes and new syndromes can be identified and clearly described together with the known mutations. In order to substantially advance the field of Mendelian disorders, the genetics community must agree to develop and utilize a repository for detailed clinical descriptions, family histories, and, ideally, any available genetic data. The main barrier to this is asking clinicians and researchers to relinquish traditional credit for their efforts. This is not to say it can't be done, but that consideration needs to be given to the ideal model for optimizing the findings for the patients and the researchers.

Identification of the genetic basis for neurodevelopmental disorders increases our understanding of how the normal brain develops and functions. Knowing the genetic cause of these disorders facilitates cellular and animal modeling. These approaches are especially important for neurological disorders where the diseased tissue is not readily available. For example, induced pluripotent stem cell (iPSC) modeling in Timothy syndrome (a rare monogenic autism syndrome) identified a compound that rescued the cellular phenotype[184]. Similarly, we have recently derived neurons from the fibroblasts of an ASNS deficiency patient; these will be characterized for phenotypic abnormalities and can be used for drug screening to rescue any observed neuronal abnormalities. In addition, we are undergoing experiments with Asns-deficient mice to alter their diet as a possible therapeutic intervention. While the ultimate goal is reversal of human neurodevelopmental disorders, drugs that effectively alleviate the symptoms are vitally needed.

As exemplified by cystic fibrosis, this process can take an extremely long time. The *CFTR* gene was identified in 1985, yet the first drug specifically acting on *CFTR* was not available until over 20 years later[185]. One major challenge with rare Mendelian diseases is that pharmaceutical companies have little incentive to take time and money to develop a drug that will only be purchased by a small number of patients. The high locus heterogeneity of epilepsy, autism, and other neurodevelopmental disorders results in the same small patient population for targeted drug therapies. However, as we

continue to identify novel genes for these neurodevelopmental disorders, it may be possible to organize all the different rare genetic risk factors into a discrete number of alternative disease-associated pathways. Once such mappings are established between collections of genetic risk factors and the pathways they affect, it will become possible to target those pathways in drug development efforts and tailor treatment for patients having different pathways perturbed.

6.1 Epilepsy future directions

How can we identify additional risk factors for the epilepsies? In the complex epilepsies, it seems conceivable that hundreds of different genes are responsible and each would explain only a very small proportion of epilepsy cases. Without massive sample sizes, obtaining sufficient evidence for genes of small effect will be nearly impossible. Additionally, the complex inheritance patterns observed in multiplex families also suggest that gene-gene (discussed in 6.2) or gene-environment interactions may play a role in epilepsy risk. Obtaining sufficiently powered cohorts to examine these questions is extremely challenging. Importantly, the next project in Epi4K is a large-scale effort to perform NGS sequencing on ~300 multiplex families (>3 affected individuals, many with non-Mendelian inheritance patterns) and another 1,500 pairs of first-degree relatives with epilepsy[60]. This work will draw from the lessons learned in my body of work, and by utilizing increased sample sizes, it will likely reveal additional risk factors for the epilepsies.

We have yet to investigate the importance of several additional classes of genetic variation in the epilepsies. In comparison to protein-coding mutations, we do not have a clear understanding of how non-coding variants should be cataloged or how they might lead to disease. Therefore, most NGS studies conducted to date have disproportionately focused on coding variants, and non-coding variants have yet to be directly and thoroughly interrogated for their association with epilepsy. It is conceivable that non-coding mutations lie in regulatory regions altering gene expression of known or novel epilepsy genes. Chemical modifications to the DNA and histones, the epigenome, alter the overall genome structure, tightly compacting inactive genes; alterations to the epigenome can also impact gene expression and may lead to disease. Epigenetic marks differ by cell type and are dynamic throughout life, responding to environmental changes, and thus studying epigenetics demands directly examining the tissue of interest. Another hypothesis is that somatic mutations, occurring in critical brain regions, explain some proportion of epilepsy cases. Testing this hypothesis would also require access to brain tissue. Finally, recent evidence has emerged supporting a possible role for a class of small (~22 nucleotides), non-coding RNA called microRNA (miRNA) in epilepsy. MiRNA molecules post-transcriptionally regulate the level of mRNA for a gene, and miRNAs can target multiple genes/proteins. It is possible that alterations to any one of these may also play a role in epilepsy susceptibility.

One central question is how we will we go from the identification of novel epilepsy genes to targeted therapies for patients. One important step in drug development is showing efficacy in animal models. Epilepsy has been studied in rodents for many years. The physiological endpoint for epilepsy, a seizure, is readily detectable in rodent models. A large number of genes are associated with seizures in mice, including both established human epilepsy genes (Table 2) and novel genes not yet associated with human phenotypes[64]. Interestingly, gene knock-outs in different genetic backgrounds (mouse strains) show variation in seizure penetrance, suggesting these mice may also be good models for complex epilepsies. In drug development, mouse models are a very powerful tool, but they can also be a rate-limiting step. One alternative is to model human mutations in cellular systems. It is now possible to rapidly and cost-effectively "edit" cells to introduce the mutation(s) of interest using RNA-guided DNA endonuclease (CRISPR) approaches[186]. Appropriate disease-modeling requires assaying relevant cell types; for epilepsy modeling, CRISPR editing can be used in iPSCs, and subsequently these cells can be differentiated into different types of neurons. The electrophysiology and overall network behavior can then be examined in these living neuronal networks by multi-electrode arrays (MEAs). To screen drugs in a mutation dependent fashion, MEAs can be used to first assess the mutant neuronal phenotypes (compared to wild type neuronal networks) and then treat the cells with drugs/compounds to screen for those that restore wild-type phenotypes. Additionally,

phenotypic comparisons amongst mutations in different genes can be used to assemble epilepsy mutations into distinct functional groups, which may respond similarly to the same drug compounds, even if the gene products themselves are not known to be in the same molecular pathway. Such an approach offers a medium-throughput method for screening drugs in a genetically informed way and may facilitate the identification of new epilepsy drugs.

The body of work presented here contributed greatly to our understanding of how to design effective genetic studies for multiplex epilepsy families. First, and probably most importantly, we need much larger sample sizes given the observed genetic heterogeneity. Second, despite the ability to readily narrow the number of candidate variants by sequencing distantly related relatives, alternate approaches paying attention to specific seizure types and overall phenotypes may be important for families with multiple segregating variants. Third, this work suggests a framework for selection of the top candidate genes by leveraging multiple layers of information. Finally, a number of candidates have been identified through this project, some of which may be confirmed with additional follow up and/or sequencing of the large number of multiplex families proposed by Epi4K.

6.2 Modifier genetics

In recent years, efforts to uncover the etiology of neurodevelopmental disorders have revealed a number of rare copy number variants (CNVs) that have variable expressivity within and across clinically distinct disorders. For example, CNVs in 1q21.1 are associated with a number of phenotypes, including intellectual disability (ID), microcephaly, schizophrenia and autism spectrum disorder (ASD)[32]. Such observations have brought the potential role of modifier mutations to the forefront of neurodevelopmental disease genetics.

Given the complexity of neurodevelopmental disease, the increasing support for major risk factors interacting strongly with one another and with the genetic background should come as no surprise. Work in model organisms has revealed rampant genetic interactions. For example, a single mutation in any of the approximately 1,000 essential genes of yeast induces lethality, but it is estimated that there are 200 times as many digenic combinations resulting in synthetic lethality[187]. In contrast, we have only a handful of examples in human diseases where multiple hits are required for manifestation of a disease. Unfortunately, identification of genetic interactions in humans has proven difficult. Even in Mendelian diseases, where the genetic architecture is simplified by low locus heterogeneity, progress has been slow, with only a few robust examples, including sickle cell anemia and cystic fibrosis[188]. In

both cases, the phenotypic expression is modified by variants outside of the disease-causing gene. In cystic fibrosis, as in many other diseases, different mutations within the disease-causing gene *CFTR* can also result in differences in disease severity[188], making it harder to identify interactions.

Recently, the genetic interactions responsible for thrombocytopenia with absent radii syndrome were investigated[189]. By focusing on patients harboring a previously associated microdeletion in 1q21.1, the researchers identified two different low-frequency variants in the regulatory region of *RBM8A*. The combination of either variant with the original microdeletion is sufficient to cause this disorder. Subsequently, patients lacking the microdeletion were found to carry novel null mutations in *RBM8A*, thus resolving the responsible gene within the 1q21.1 region. This compound inheritance mechanism explained 53 of 55 cases ($P < 5 \times 10^{-228}$) and provides a simplified model that can be applied in studies of neurodevelopmental disorders. However, given the low frequency of known risk alleles and the extreme genetic heterogeneity, identifying well-powered cohorts of genetically homogeneous samples will be no small task.

Common complex diseases represent a particular challenge for studying modifier genetics. Perhaps the most fundamental constraint is that the high locus heterogeneity complicates identification of patients with similar 'primary' mutations in order to ask how these interact with modifiers. So far, such efforts have been modest. In the case of heterozygous microdeletions, one possibility is that variable expressivity is

due to newly hemizygous deleterious mutations in distinct genes on the remaining chromosome. This possibility has been tested in only a few studies to date [30,32], with no clear evidence of genetic modifiers on the intact chromosome. Admittedly, the small sample sizes in these studies mean that it is difficult to rule out this possibility even for the deletion regions that have been tested. Other studies have looked elsewhere in the genome for evidence of genetic interactions. Girirajan *et al.* [190] performed a genome-wide scan for CNVs in ID patients carrying the 16p12.1 microdeletion and identified a non-specific enrichment of large CNVs that correlated with a more severe clinical phenotype. Perhaps the presence of one causal CNV more readily allows the presence of others compared with controls. It remains unclear whether these observations are reflective of a primary driver with secondary modifiers or if some combination of multiple hits is necessary for manifestation of the phenotype.

Developing well-powered modifier studies for epilepsy and other neurodevelopmental disorders will be a very significant challenge. Either way, to obtain the statistical evidence needed to prove such associations, large, well-phenotyped and homogeneous cohorts must be compiled, again highlighting the need for collaborative efforts among researchers. On a broader level, identification of multiple genetic aberrations in patients, or unaffected individuals, may reveal important combinatorial effects on phenotypic variability and novel underlying biological interactions.

6.3 Concluding remarks

The studied neurodevelopmental disorders exhibited a range of genetic complexity, from clear Mendelian disorders to common complex disorders, resulting in varying degrees of success in identifying the underlying causal genetic variants using NGS. These mixed outcomes will likely to be observed outside of neurodevelopmental disorders and human geneticists now face the problem of having not one clear path forward, but several critical future research directions.

In ASNS deficiency, we have definitively identified the causal mutations which all lie within the *ASNS* gene. However, despite elucidating the genetic etiology of this disorder, much effort will be required to understand the molecular mechanisms by which mutations in this gene lead to the clinical presentation. We also do not yet understand what factors are responsible for variation in clinical presentation. The research presented here has outlined a number of first steps towards unraveling the molecular mechanisms, but it is important to recognize that this work is rather open ended and that the biological discoveries along the way will guide future experiments. Such open ended molecular experimentation, is a critical issue across many Mendelian diseases recently solved by NGS. While it is exciting that the genetics community is rapidly uncovering novel disease genes, many more molecular biology experiments are needed to inform our understanding of the molecular pathology of these disorders and their direct connection to the clinical phenotypes.

Sometimes application of NGS to these neurodevelopmental disorders resulted only in a strong candidate variant(s)/gene(s) that we are unable to prove at this time. For example, in addition to the 32% of previously undiagnosed disorder patients where a genetic diagnosis was successfully made after NGS, in the remaining undiagnosed patients, ~30% have novel candidate variants, some of which will likely prove real when additional patients with these phenotypes are recognized. The problem is that we might not always be able to obtain additional patients with overlapping phenotypes. Again, this issue would be greatly furthered by increased collaboration amongst clinicians and researchers, especially when only good candidates are identified and thus not independently publishable. For diseases with high locus heterogeneity, even if we obtain additional well-phenotyped samples, we may still not have enough genetic homogeneity amongst these samples to prove these good candidate variants right away. Therefore, we need to develop additional frameworks for recognizing novel variants and improve our understanding of noncoding variation and its role in disease.

Finally, we also observe some disorders, such as the epileptic encephalopathies, for which many different causal mutations are known but it is not clear if or how we should group these patients. It appears that at least some of the responsible genes will congregate on the same molecular pathway, and if these genes converged into a discrete number of alternative disease-associated pathways this would greatly facilitate drug development efforts and tailored treatment options for patients having different

pathways perturbed. However, we might also be able to group patients by those where the underlying mutations result in similar molecular defects despite lack of known direct interactions between their gene products. Unraveling the latter possibility will also require more molecular approaches, such as neuronal network phenotyping with MEAs for the epilepsies.

Collectively, this body of work has securely implicated three novel neurodevelopmental disease genes that inform the underlying pathology of these disorders. In three of these four studies, this work has highlighted additional candidate variants and genes that may ultimately be validated as disease-causing as sample sizes increase and novel analysis frameworks are developed. The future of human genetics now depends on our ability to extend genetic discoveries to expose the molecular mechanisms of pathogenicity that can explain patient phenotypes, our ability to identify novel pathogenic variants with improved analysis frameworks, our ability to collaboratively increase sample sizes and compile detailed phenotypic information, and finally, our ability to deconvolute complex phenotypes by accurately categorizing patients based on similar genetic/molecular etiologies.

Appendix A: Additional phenotypic information for ASNS deficiency patients

A.1 Comparison to primary microcephaly (MCPH)

A.1.1 Primary microcephaly (MCPH) and MCPH genes

Despite congenital microcephaly being a consistent feature of this syndrome, our patients do not fit the definition of primary microcephaly (MCPH). MCPH disorders are characterized by: a profoundly small head at birth, modest developmental delay (considering the dramatic microcephaly), mild or no seizures, and little or no deterioration. To date, there are seven distinct chromosomal loci for autosomal recessive primary microcephaly: MCPH1-MCPH7 (Online Mendelian Inheritance in Man (OMIM; http://www.omim.org; MIM# 251200, 604317, 604804, 604321, 608716, 608393, and 612703, respectively). The causal genes have been identified for all of these loci (Table 31), including the recent discoveries of causal genes in MCPH2 [191,192] and MCPH4 [193]. Multiple types of casual mutations (e.g., non-synonymous, small deletions) have been identified in each of these genes [191-194]. The observed genetic heterogeneity is not surprising given the broad clinical spectrum of primary microcephaly [195].

Table 31. Seven primary microcephaly loci.

Locus	Location	Gene	Coordinates (NCBI36)	MIM
MCPH1	8p23	Microcephalin (MCPH1)	chr8:6,251,529-6,493,434	607117
MCPH2	19q13	WD repeat-containing protein 62 (WDR62)	chr19:41,237,623-41,245,393	613583
MCPH3	9q33	CDK5 regulatory subunit-associated protein 2 (CDK5RAP2)	chr9:122,190,968-122,250,225	608201
MCPH4	15q21	Centrosomal protein, 152-KD (CEP152)	chr15:37,900,001-42,700,000	613529
MCPH5	1q31	Abnormal spindle-like, microcephaly-associated (ASPM)	chr1:195,319,997-195,382,287	605481
MCPH6	13q12	Centromeric protein J (CENPJ)	chr13:24,354,412-24,395,085	609279
MCPH7	1p32	SCL/TAL1-interrupting locus (STIL)	chr1:47,488,398-47,552,406	181590

Table 32. Predicted homozygous regions overlapping primary microcephaly loci.

Individual	Homozygous region start coordinate	Overlap gene	Overlap start coordinate	Overlap stop coordinate	Size of overlap (KB)
Family A (II-1)	chr13:24181657	CENPJ	24354412	24395085	40.674
Family B (II-4)	chr8:6259807	MCPH1	6259807	6364841	105.035
Family A (II-1)	chr8:6289591	MCPH1	6289591	6466364	176.774
Family A (II-1)	chr9:122210554	CDK5RAP2	122210554	122250225	39.672

A.1.2 MCPH genes and homozygosity mapping in ASNS deficiency patients

Despite not matching the MCPH phenotype, we still wanted to exclude the possible involvement of previously identified primary microcephaly genes (Table 31), these loci were analyzed with respect to the 1,532 homozygous regions identified in any of the three original case genomes (families A and B). There were four homozygous regions that overlapped one of the microcephaly-associated genes, however these regions were never homozygous in more than one patient (Table 32).

Parenthetically, exome sequencing has the potential to miss casual variants if the exon harboring it is not captured or if there is undetectable structural variation. Our analysis, however, suggests that homozygous variants on a single haplotype have not been missed in any of the microcephaly genes.

A.1.3 ASNS deficiency and other neurometabolic disorders

The new syndrome most resembles a host of neurometabolic disorders [196-198] that show microcephaly at birth and progressive deterioration, including progressive brain atrophy. Amish microcephaly (MCPHA, OMIM #607196) is one of the most well-defined examples of a metabolic disorder which resembles the new syndrome. MCPHA is characterized by infantile microcephaly and α-ketoglutaric aciduria. However, we noted that Amish microcephaly is more severe than the new syndrome in that Amish microcephaly is typically diagnosed at the second trimester (compared to the third

trimester in our patients) and is associated with arrest of brain development *in utero* [196,197].

A.2 Families A & B

Patients were born at term with a small head circumference (at least two standard deviations below average). From early infancy they showed feeding difficulties, jitteriness, and increased disabling spastic tone, followed by evolution of intractable infantile spasms around the age of three months. The spasms were replaced by frequent myoclonic and tonic seizures, partially responsive to anticonvulsant therapy. By the age of one year, the children failed to reach significant milestones, showed spastic quadriparesis, profound ID, and cortical blindness and had a head circumference seven standard deviations below average. Physical examination was remarkable for mild dysmorphic features including receding forehead, big fleshy ears and relatively big hands and feet. Electroencephalogram (EEG) evolved from burst suppression pattern at three months to multiple independent spike foci (MISF) on a slow background. Visual and auditory evoked potentials (VEP, BERA), as well as electroretinogram (ERG), were normal. Brain CT at two months of age showed cortical atrophy without calcifications. Magnetic resonance imaging (MRI) at the age of three months showed diffuse cerebral atrophy, enlarged extra-ventricular spaces with normal size cerebellum and brainstem and at the age of 18 months, progressive cortical and subcortical atrophy with delayed myelination, moderately enlarged lateral ventricles

and thin corpus callosum. The brain stem was small, while the cerebellum was only very mildly affected (Figure 43).

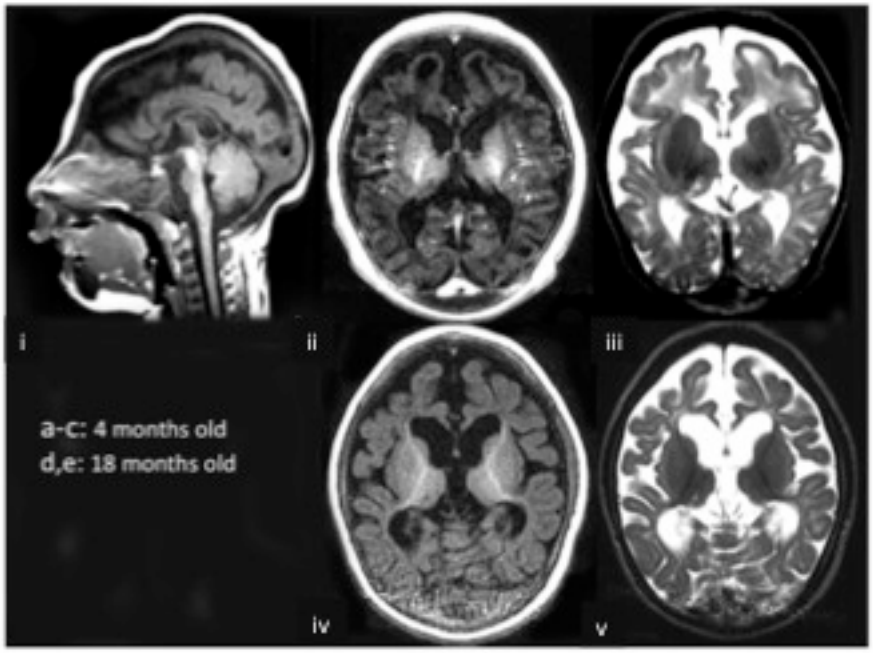


Figure 43. Brain MRI from patient A.II.1.

MRI from individual A.II.1 in family A. T1W sagittal image **(i)** at 4 months of age demonstrates micrognathia, sloping forehead, thin brainstem and relatively normal cerebellum. T1 axial image **(ii)** reveals severe cortical and subcortical atrophy and delay in myelin maturation, with myelin only in the posterior limbs of internal capsule. T2 axial image **(iii)** reveals abnormally increased signal intensity of the white matter for age, prominent ventricles and pericerebral spaces over the frontal convexity. T1W axial image **(iv)** at 18 months of age confirms persistent myelin delay and parenchymal volume loss. Concurrent T2W axial image **(v)** shows abnormal signal and atrophy of the white matter, progressive ventriculomegaly and pericerebral spaces due to volume loss.

In each of the families one pregnancy was terminated due to arrest of head growth detected by a fetal ultrasound. In the aborted child from family A, a fetal brain

MRI was also performed, which identified bilateral cortical dysplasia and cerebellar hypoplasia, both of which were confirmed by autopsy.

A.3 Family C

The patients were three siblings born to a consanguineous Bangladeshi couple in good health. An older sister is unaffected. The parents are second cousins in that the mother's maternal grandmother and the father's maternal grandmother were siblings.

Patient C.II.1. The couple's first child, a male, was born at term following an uncomplicated pregnancy. Delivery was by C-section due to thick meconium as well as fetal distress. His APGAR score were 5 and 8 at 1 and 5 minutes respectively, his birth weight was 3400 g (50th centile) and his head circumference (OFC) was 30.5 cm (-3.5SD). He was noted to be jittery during the first hour of life with episodes of hyperekplexia. He required oxygen at birth but quickly was weaned. EEG recording at 9 days of age was moderately abnormal due to disorganized and discontinuous background. No epileptiform discharges were seen. Initial MRI showed left transverse sinus thrombosis and cerebral dysgenesis and renal ultrasound showed echogenic kidneys. He was treated with anticoagulants and a CT scan venography at 2 months of age showed that the thrombus cleared. He had no developmental milestones and showed feeding difficulties with episodes of aspirations for which a G-tube was inserted. Repeat brain MRI revealed an immature brain with delayed myelination as well as a left transverse sinus thrombosis. The infant continued having tremor, persistent hyperekplexia and was

discharged from the hospital at a month of age only to be re-admitted at 3.5 months of age with respiratory failure. He was intubated and ventilated but multiple attempts at extubation over the course of 3.5 weeks failed due to central respiratory drive failure. Given the infant's futile outlook, the infant life support was withdrawn at 4 months of age. Investigation including chromosome analysis, metabolic work-up and studies for congenital infections showed no abnormalities. The family declined both autopsy and DNA banking.

The couple's second pregnancy resulted in a healthy daughter following an uneventful pregnancy and delivery. At 8 years of age she has no health or developmental issues.

Patient C.II.3. The couple's third pregnancy was followed closely and chorionic villi sampling (CVS) for maternal age revealed a normal male karyotype (46,XY). Serial ultrasounds at 19, 22, 26, and 32 weeks gestation all revealed normal head size at each gestational age. Delivery was by repeat caesarean section at 39 weeks gestation. Birthweight was 3520g (50th centile) and OFC was 33 cm (-1SD/-2SD). Shortly after birth he developed jittery seizure-like movements and hyperekplexia mirroring those of his late brother. Extensive investigation including plasma amino acids [asparagine value was 12 mmol/L after birth and 17 mmol/L at 2 months (normal 16-21 mmol/L), urinary organic acids, plasma very long chain fatty acids, 7 dehydrocholesterol, transferrin isoelectric focusing, free and total carnitine, serum ceruloplasmin, blood CK, urine

mucopolysaccharides and urine oligosaccharides were normal. Chromosome microarray analysis was normal. Initial EEG demonstrated intermittent right and left frontal sharp waves occurring independently of uncertain significance, possibly indicative of underlying cerebral dysfunction. No definitive seizure activity was noted. A repeat EEG done at a month of age was contaminated by a large amount of movement artifact. Within these limitations, the waking background activity did not show any definite abnormality, nor did a brief period of active sleep. Although not definitely abnormal, there were very frequent interictal sharp waves recorded independently over both frontopolar regions with a definite left sided predominance. No ictal epileptiform discharges were recorded. A repeat EEG at 2 months of age was abnormal with discontinuous and asynchronous background and positive Rolandic sharp waves seen over the left more than right hemisphere. No electrographic seizures were seen. Basal auditory evoked potential was normal, visual evoked potential and electroretinogram showed delayed cortical response by 45 msec and somatosensory evoked potential of the median nerve showed absent cortical response bilaterally. EMG was normal. Brain MRIs revealed volume loss of cerebral hemispheres and pons, with normal appearing cerebellum, delayed myelination of posterior limbs of internal capsule and dorsal brainstem, with mild restricted diffusion of both thalami and possibly basal ganglia. There was mildly elevated lactate peak on MRS and no imaging evidence of prenatal asphyxia. The findings were similar to the MRI findings performed on his late brother.

At 2 months of age he was noted to have right hemidiaphragm palsy and had increasing respiratory failure. He developed hypothermia and hypercapnia and was removed from respiratory support at 3 months of age.

Patient C.II.4. The couple's fourth pregnancy was again followed closely and CVS for maternal age revealed a normal male karyotype (46,XY). Fetal ultrasounds at 19, 28, and 30 weeks showed normal head growth. At 32 weeks a slight lag in OFC was noted, with progressive development of microcephaly seen at 32, 34 and 35 weeks gestation by ultrasound. The baby was born at 35.7 weeks gestation via repeat C-section. The APGAR scores were 9 and 9 at 1 and 5 minutes respectively. Birth weight was 3230 g (50th centile), birth length was 51.5 cm (50-97 centile) and OFC was 32 cm (-2SD). At 6 hours after birth, he was noted to have abnormal movements with jitteriness and hyperekplexia. He was initially intubated but extubated at 2 days of age and was noted to have stridor. He continued having seizure episodes with back arching and apnea and was treated with clonazepam and phenobarbital. EEG on his first day showed discontinuous background with longer interburst interval than appropriate for age with occasional positive and negative sharp waves over the left central and temporal head regions. There were occasional runs of theta and alpha rhythm of unknown significance. No seizures were captured during this recording. Brain MRI showed microcephaly, delayed myelination and under development of the brain with brainstem hypoplasia. The findings were similar to the MRIs done on the previous siblings. Plasma amino

acids and urine organic acids were normal. The anticonvulsive treatment resulted in decrease in his jitteriness and he did not have respiratory problems apart from stridor, which was attributed to laryngomalacia. He was orally fed with frequent vomiting and continued having tremor and jitteriness although this decreased with time. He was transferred to a palliative care facility and died at 6 months of age.

A.4 Family D

The patients were three brothers born to a nonconsanguineous French-Canadian couple in good health. An older sister is unaffected.

Patient D.II.1. The pregnancy and neonatal period were considered normal but on the first routine home visit by a nurse at 5 days of age, he was found to have irregular breathing, rectal temperature of 34.6°, and to have poor suction at feeds. He was examined in a local center and transferred to CHU Sainte-Justine where he was intubated and became respirator-dependent. He was microcephalic (31.5 cm; -2.5SD) with central hypotonia and increased tone in the legs and tremulous movements of the extremities. Visceral examination was normal. Glycemia, blood pH, transaminases, renal function were normal. Fasting plasma amino acids showed asparagine 11 μmol/L (normal range, 31 to 56). Urine organic acids and plasma very long chain fatty acids were interpreted as normal. Transferrin isofocalization was normal. MECP2 gene sequencing and deletion analysis were normal. Cerebral magnetic resonance imaging (MRI) showed microcephaly with reduced gyration. He was extubated and died rapidly

thereafter of respiratory failure, aged 9 days. Neuropathological examination revealed microcephaly, although the gyral pattern was preserved grossly on external examination. There was cortical dysgenesis (nests of cells in the subpial region and multiple thin layers of nuclei in the cortex, thinner and more numerous than normal). Also present were absence of the bulbar arcuate nucleus, marked periventricular leukomalacia; and marked gliosis. The visceral examination was normal.

Patient D.II.2. The patient was born after a normal pregnancy at 39 5/7 weeks by spontaneous vaginal delivery. Fetal MRI at 31.5 weeks gestation and fetal ultrasound at 35 weeks, were normal. On neonatal examination, jerky and tremulous movements and limb hypertonia were noted. Body temperature was repeatedly normal. Head circumference was 31 cm (-3SD). At 4 days of age an EEG showed deficient electrical activity with bursts of activity with spike activity maximal in the right fronto-centro-temporal regions. Cerebral MRI at 4 days of age revealed a reduction of the normal gyral pattern (Figure 44). The opercularisation of the sylvian fissure corresponded to that of a 34 week old child (at a chronological age of 39 weeks). There was a diffuse hyposignal on T1 imaging, demonstrating a slowness of myelination. The parieto-occipital horns were prominent, especially on the left. The subarachnoid space was prominent.

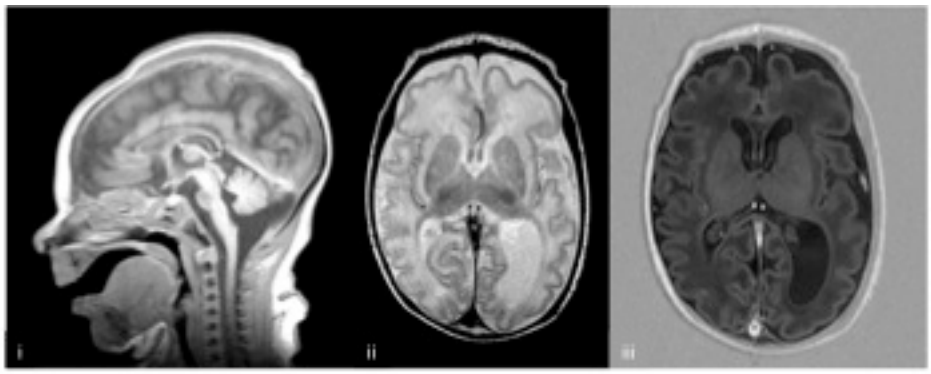


Figure 44. Brain MRI from patient D.II.2.
(i) Sagittal T1 image showing small pons. **(ii)** Axial T2 image showing decreased brain volume, delayed myelin maturation, and simplified gyral pattern predominantly in the frontal regions. **(iii)** Axial T1 image confirming the delayed myelination and the other findings seen on the Axial T2 image.

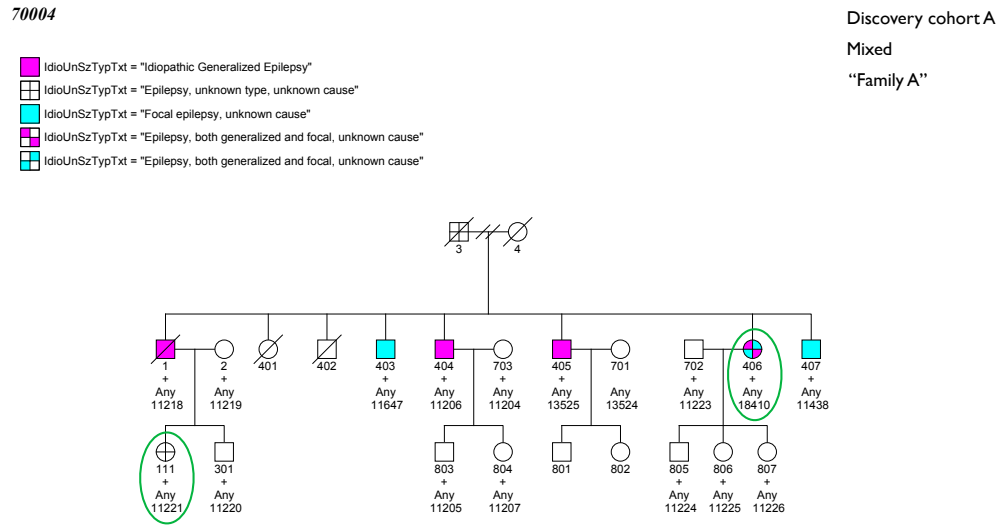
Plasma amino acids were normal including asparagine (55 μmol/L; normal, 31-56), as were urine organic acids, transferrin isofocusing, analysis of subtelomeres and ARX gene sequencing. Ophthalmologic consultation revealed cortical blindness. At age 9 months, partial motor convulsions were reported and focal epileptic activity was seen on EEG; control was achieved with carbamazepine. Examination at 10 months showed a nondysmorphic child with an absence of motor development and cranial circumference of 40 cm (<3%). He reacted to noise but showed no ocular pursuit. Divergent strabismus, axial hypotonia and limb hypertonia and hyperreflexia were present. The lower limbs were more severely affected than the upper. He died at age 11 months following an upper respiratory infection associated with respiratory acidosis (pCO2, 120; pH, 7.03) and hypothermia (30° C). At autopsy, neuropathology confirmed microcephaly with severe mesial temporal sclerosis and neuronal loss in regions CA3, CA4 and CA1,

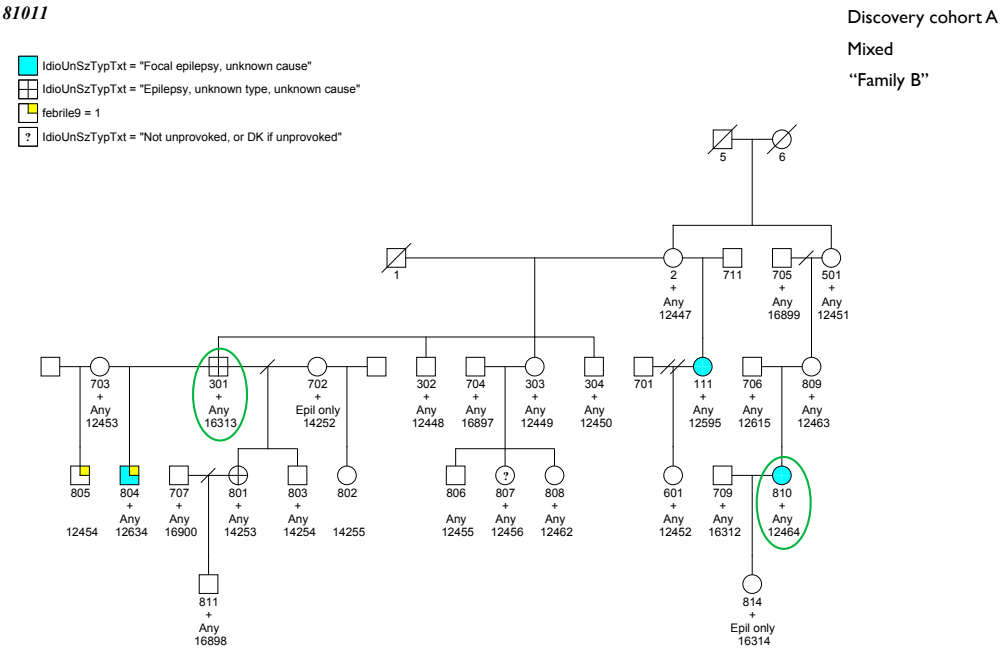
dysplasia of the olivary nucleus. The spinal cord showed hydromelia with secondary degeneration of motor neurons and reactive gliosis.

Patient D.II.3. The pregnancy was initially normal. Fetal ultrasound was normal at 25 weeks of gestation. Placental abruption with fetal distress occurred at 33 weeks, and emergency cesarean section was performed. Apgar scores were 4 at one minute, 7 at 5 minutes and 10 at 10 minutes. Birth weight was 2160 g (50-75%), length 46 cm (75%) and head circumference 30 cm (25th centile). Although the patient initially cried vigorously, he showed hypoventilation within hours of birth. Respiratory distress syndrome was diagnosed clinically and on thoracic radiographs. He was intubated. Because of apneas he was treated with caffeine on day two. After resolution of his pulmonary disease, attempts to wean him from continuous positive airway pressure failed because of the development of apnea, cyanosis and bradycardia. Tremulous movements of the extremities were noted. He was transferred to CHU Sainte-Justine on day 6. Convulsions were noted on day 8. He was extubated at 3 weeks of age. At that time he could ingest adequate amounts of formula orally. His head circumference at one month of age was 30.2 cm and at 10 months, 39.5 cm (<3%). His subsequent clinical course was similar to that of patient 2. His convulsions were controlled by small doses of Phenobarbital. He died at age 12 months of respiratory insufficiency secondary to a respiratory infection, associated with hypothermia.

Appendix B: Multiplex epilepsy family pedigrees[1]

B.1 Multiplex epilepsy families from Discovery Cohort A (n=29).





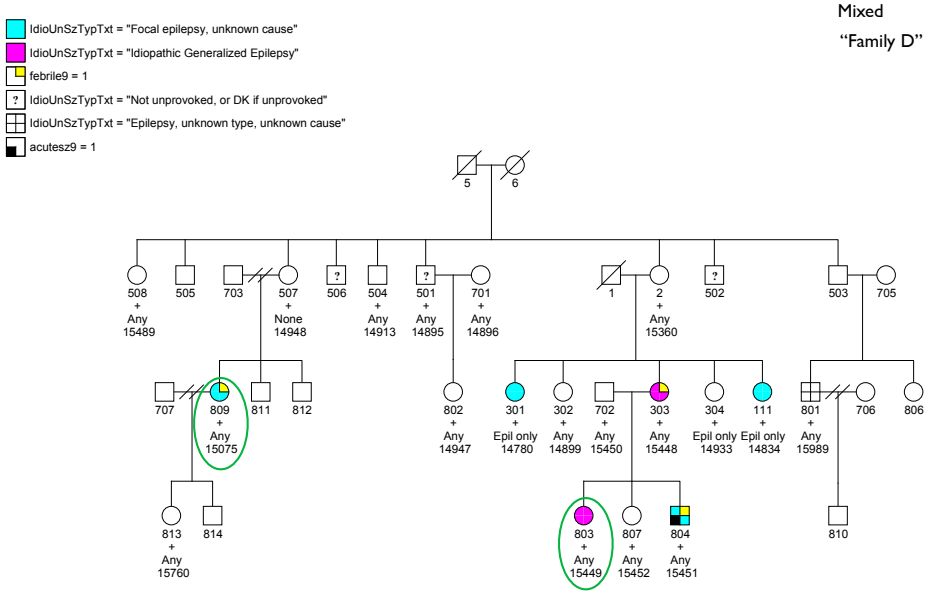
[1] Circled individuals were whole-genome or exome sequenced.

81023

Discovery cohort A

Mixed

"Family D"

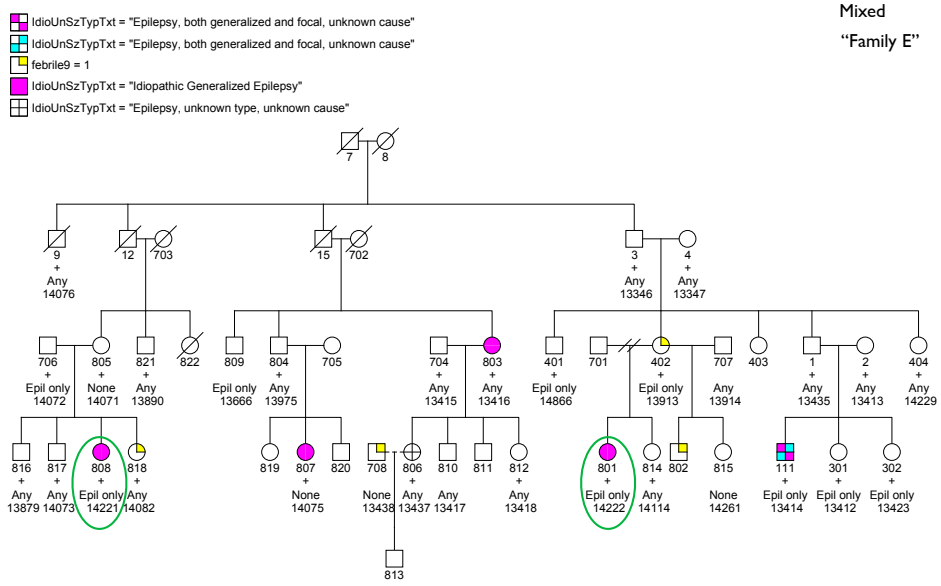


82055

Discovery cohort A

Mixed

"Family E"



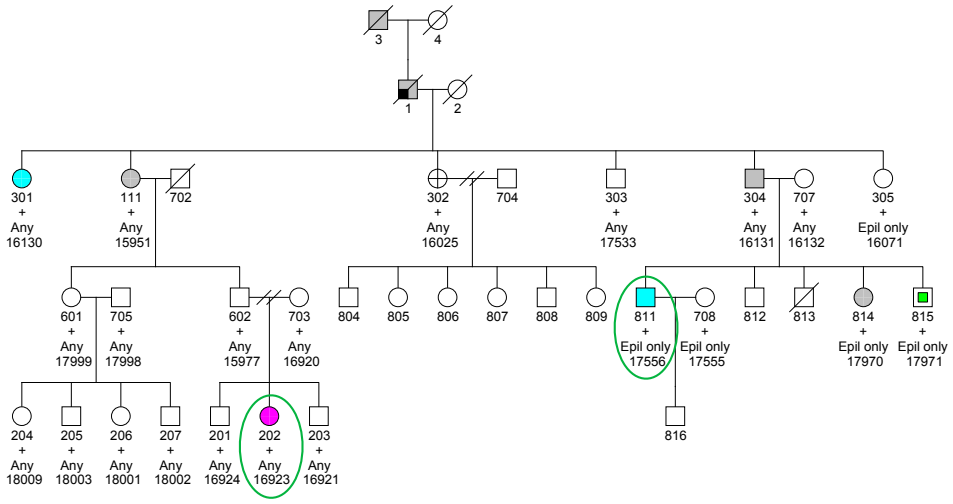
84042

Discovery cohort A

Mixed

"Family F"

IdioUnSzTypTxt = "Unprovoked seizure(s), structural/metabolic cause"
acutesz9 = 1
IdioUnSzTypTxt = "Idiopathic Generalized Epilepsy"
IdioUnSzTypTxt = "Focal epilepsy, unknown cause"
IdioUnSzTypTxt = "Epilepsy, unknown type, unknown cause"
IdioUnSzTypTxt = "Isolated unprovoked seizure, unknown type, unknown"



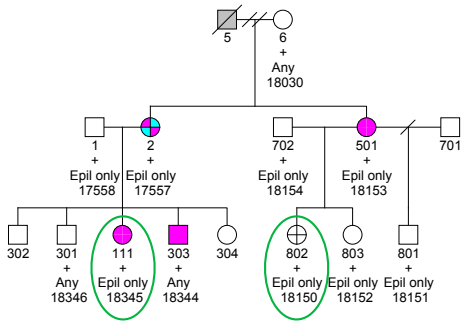
86010

Discovery cohort A

Mixed

"Family H"

IdioUnSzTypTxt = "Epilepsy, both generalized and focal, unknown cause"
IdioUnSzTypTxt = "Epilepsy, both generalized and focal, unknown cause"
IdioUnSzTypTxt = "Unprovoked seizure(s), structural/metabolic cause"
IdioUnSzTypTxt = "Idiopathic Generalized Epilepsy"
IdioUnSzTypTxt = "Epilepsy, unknown type, unknown cause"

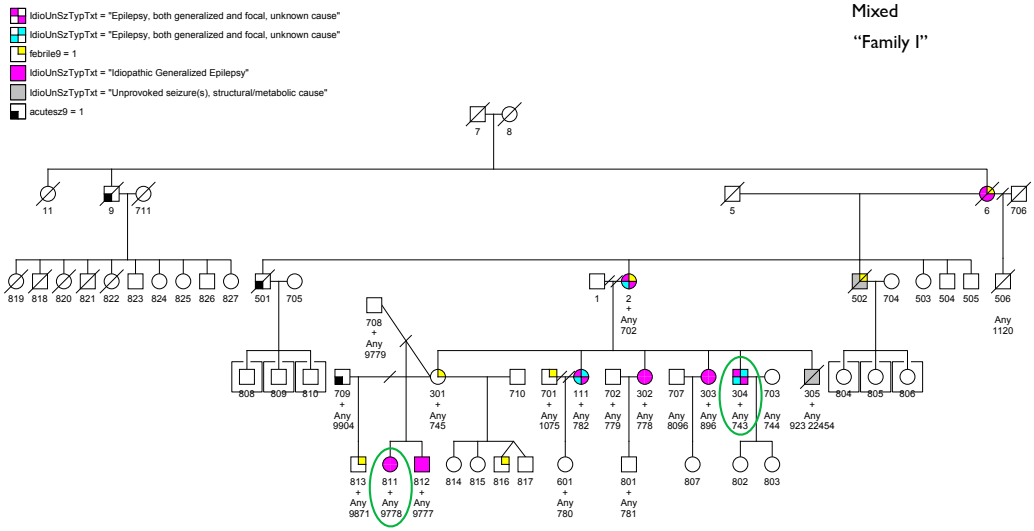


50031

Discovery cohort A

Mixed

"Family I"

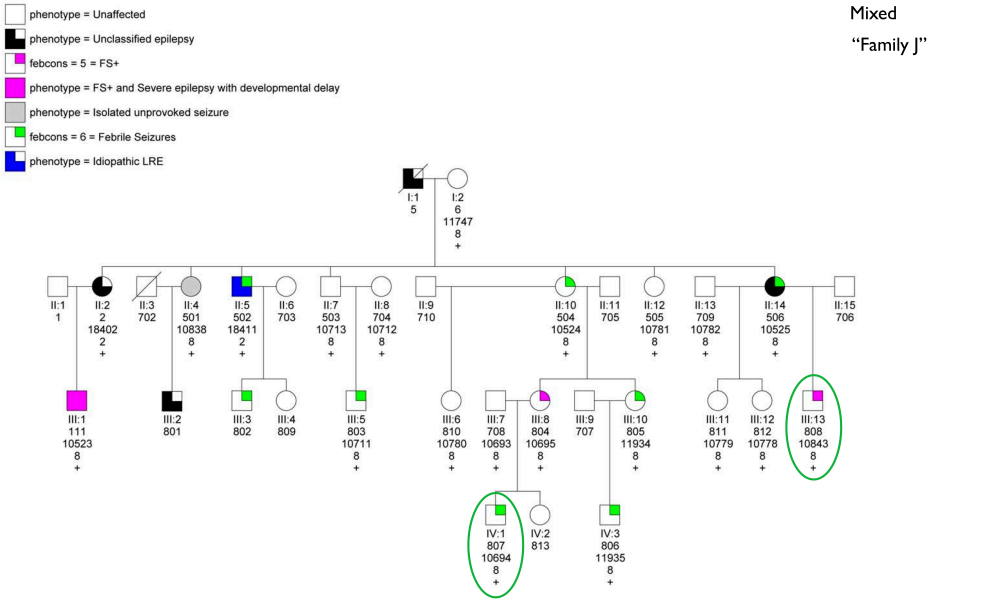


80011

Discovery cohort A

Mixed

"Family J"

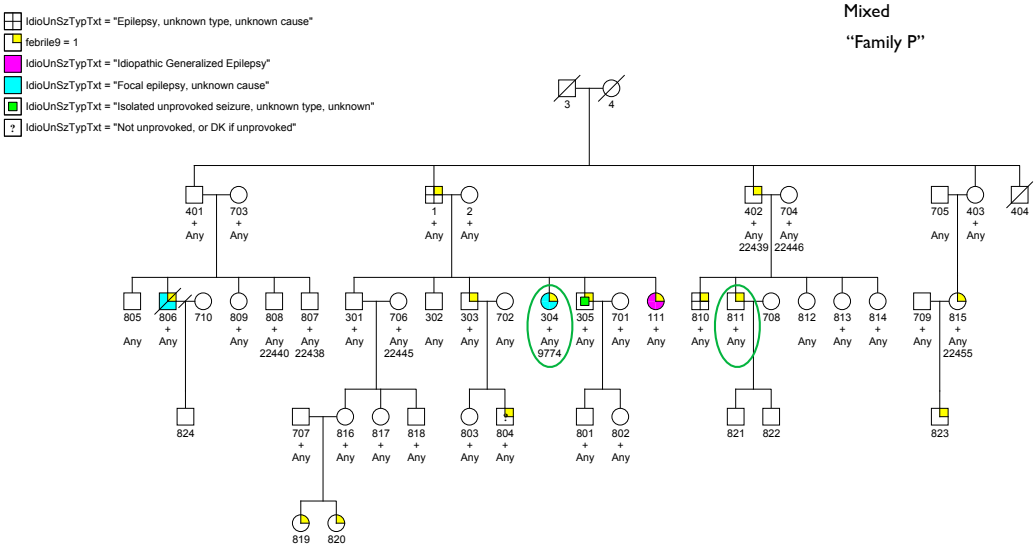


4471

Discovery cohort A

Mixed

"Family P"

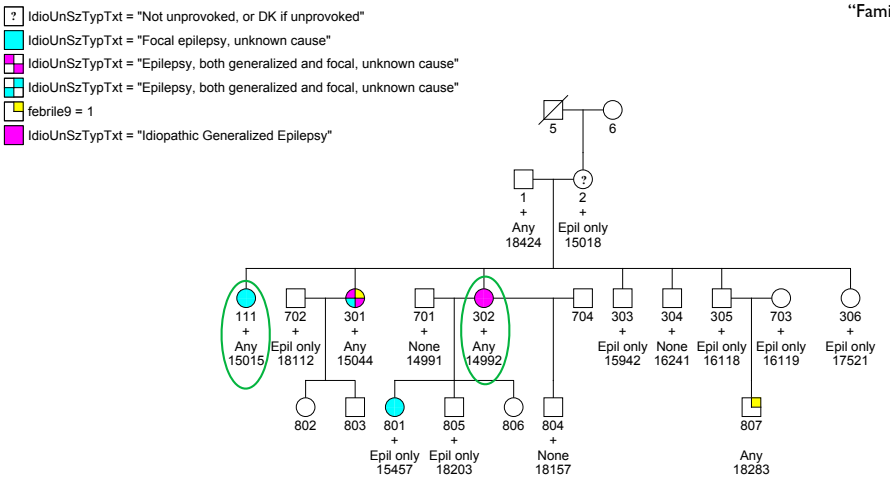


82078

Discovery cohort A

Mixed

"Family Q"

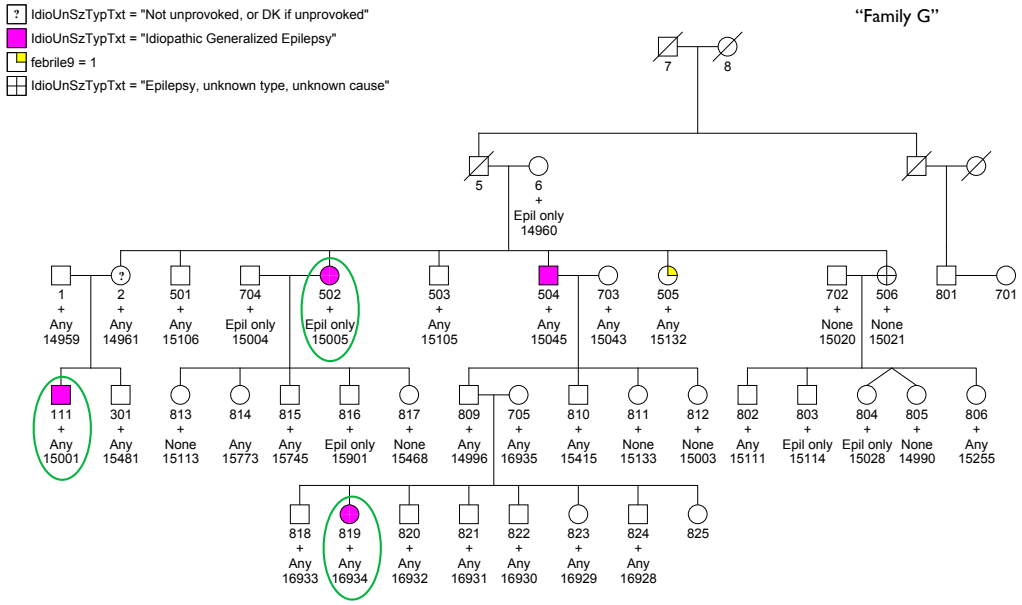


85033

Discovery cohort A

GGE

"Family G"

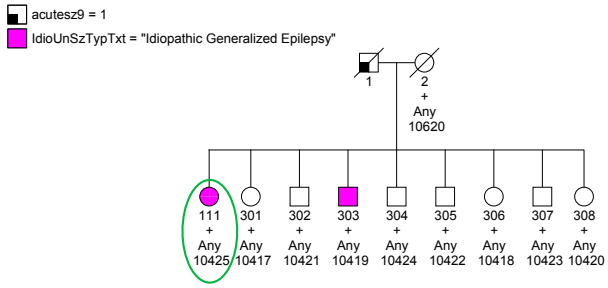


4002

Discovery cohort A

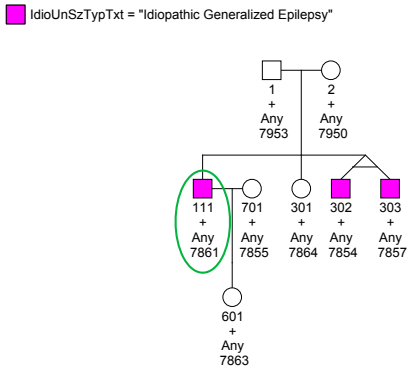
GGE

"Family T"



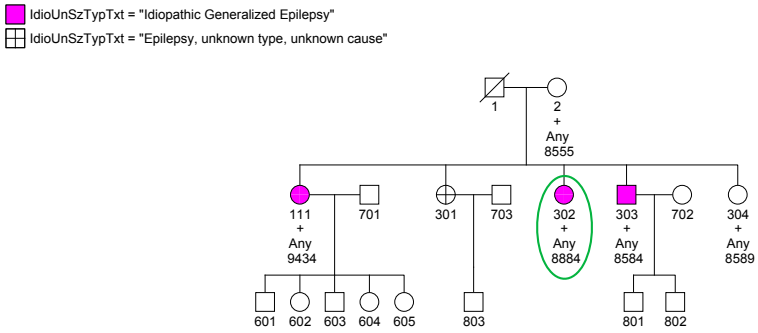
4541

Discovery cohort A
GGE
"Family U"



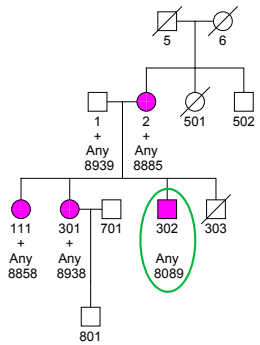
8801

Discovery cohort A
GGE
"Family V"



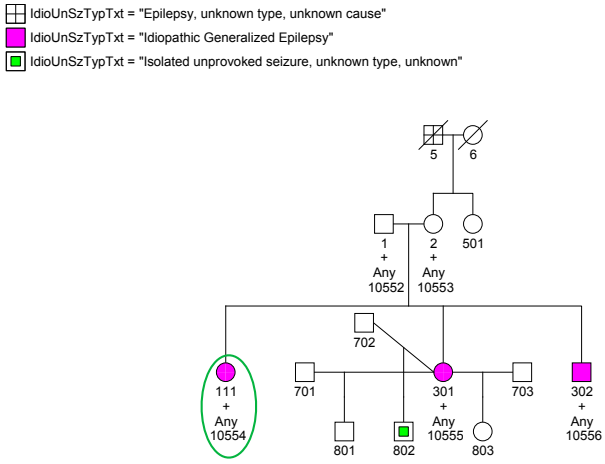
9943

Discovery cohort A
GGE
"Family W"



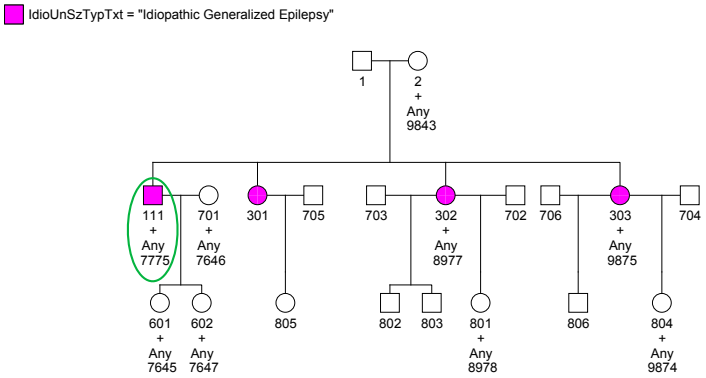
60160

Discovery cohort A
GGE
"Family X"



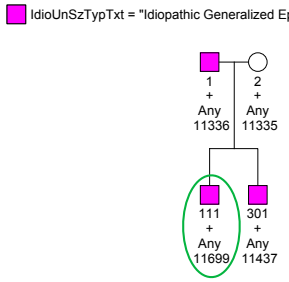
60266

Discovery cohort A
GGE
"Family Y"



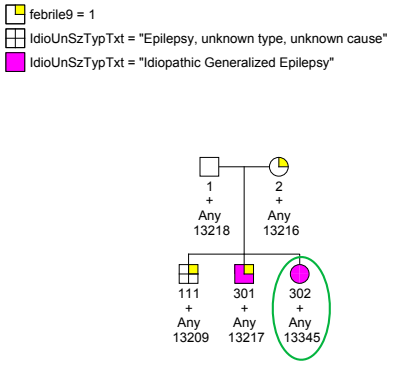
82009

Discovery cohort A
GGE
"Family Z"



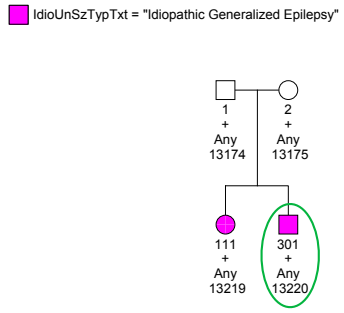
82060

Discovery cohort A
GGE
"Family AA"



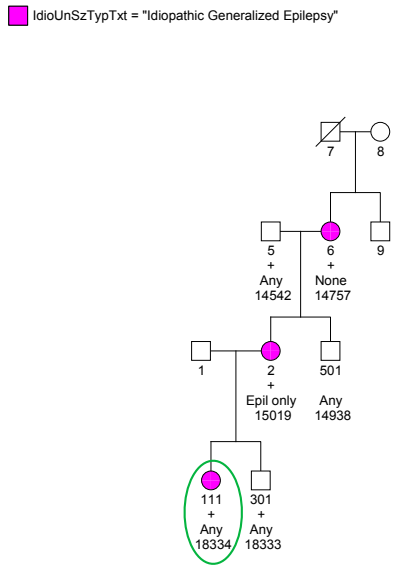
82061

Discovery cohort A
GGE
"Family BB"



82074

Discovery cohort A
GGE
"Family CC"

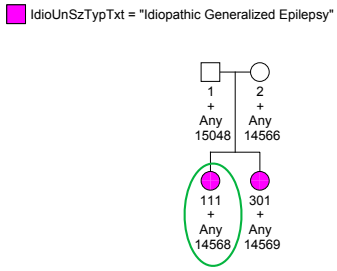


84031

Discovery cohort A

GGE

"Family DD"

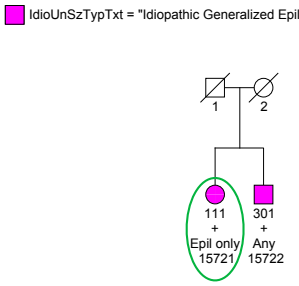


87002

Discovery cohort A

GGE

"Family EE"

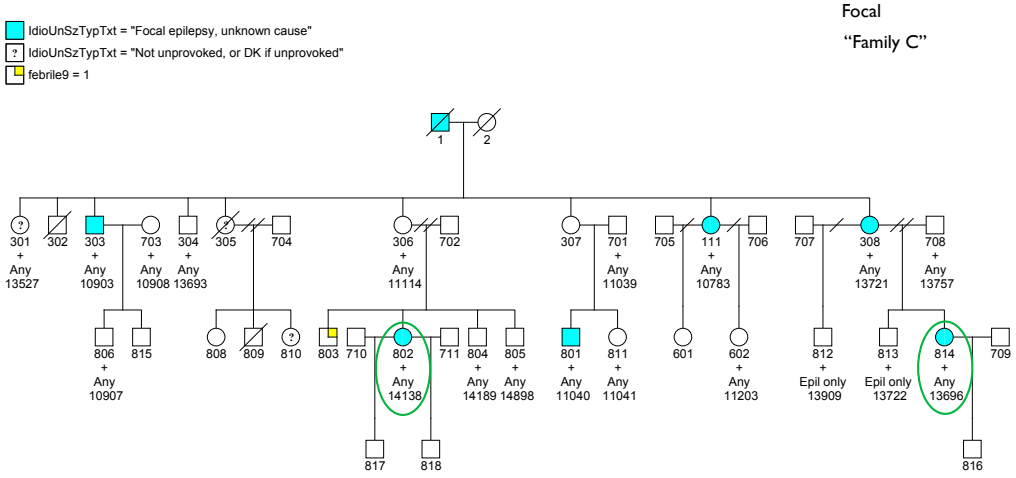


80020

Discovery cohort A

Focal

"Family C"

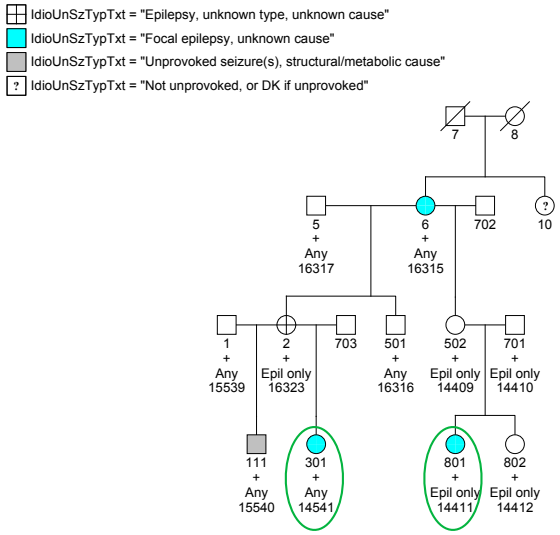


85030

Discovery cohort A

Focal

"Family K"

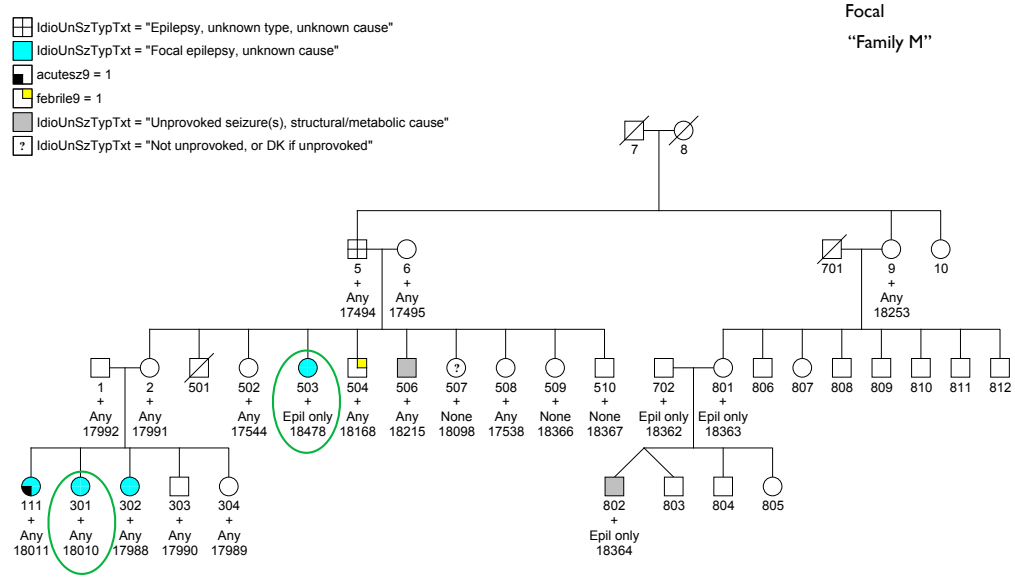


86009

Discovery cohort A

Focal

"Family M"

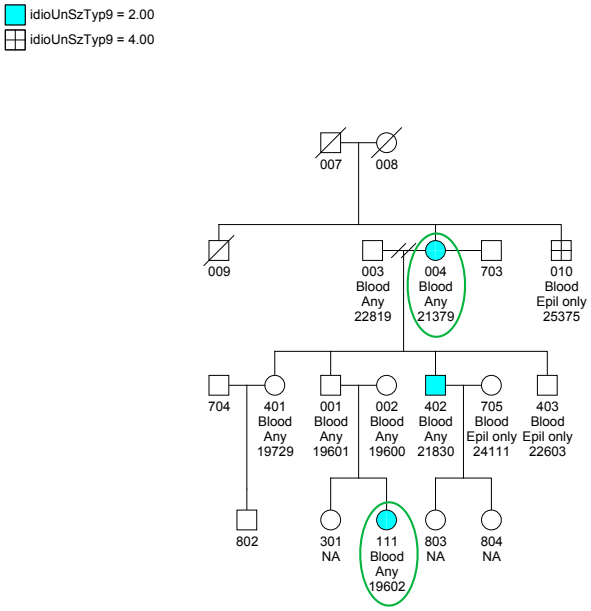


A0017

Discovery cohort A

Focal

"Family N"

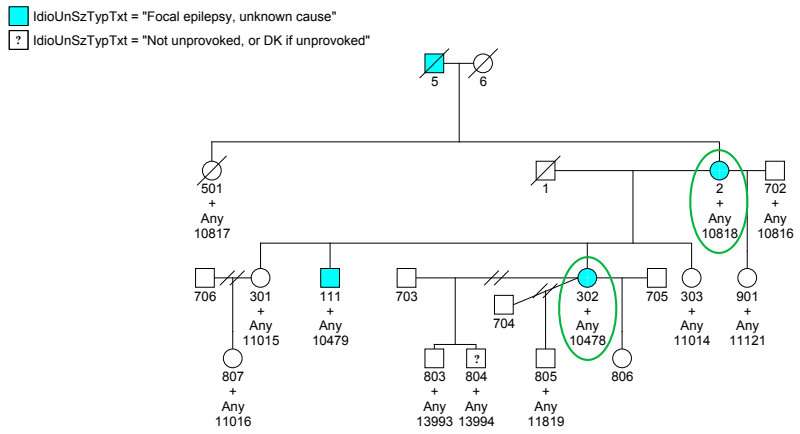


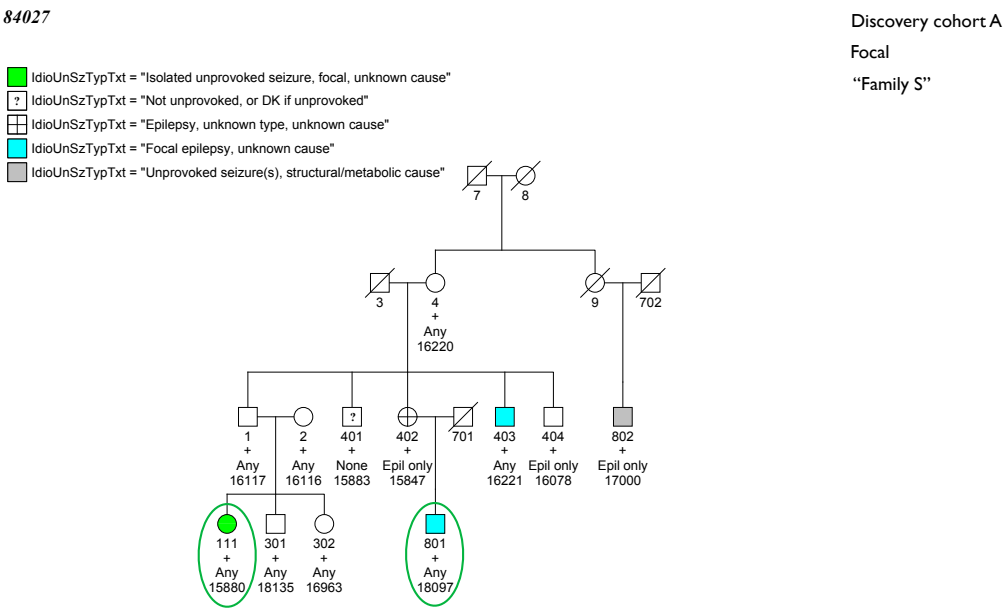
1521

Discovery cohort A

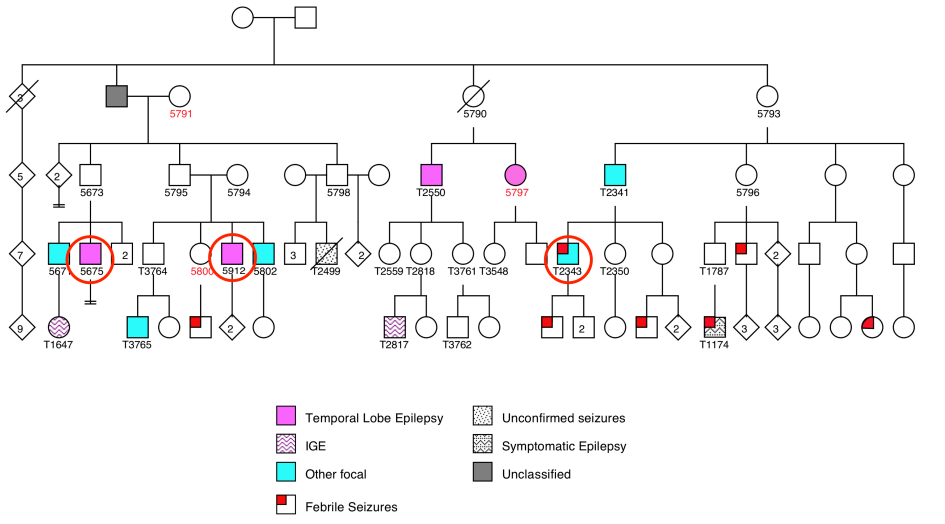
Focal

"Family R"



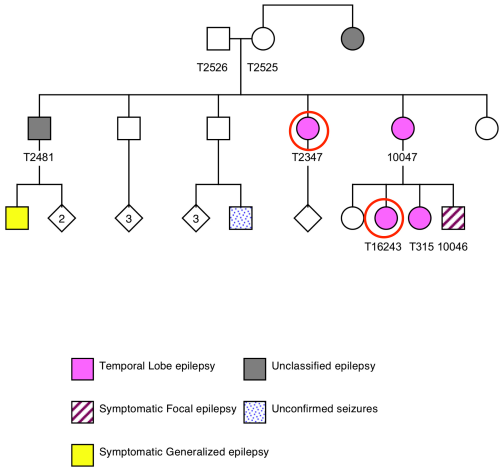


B.2 Multiplex epilepsy families from Discovery Cohort B (n=10).



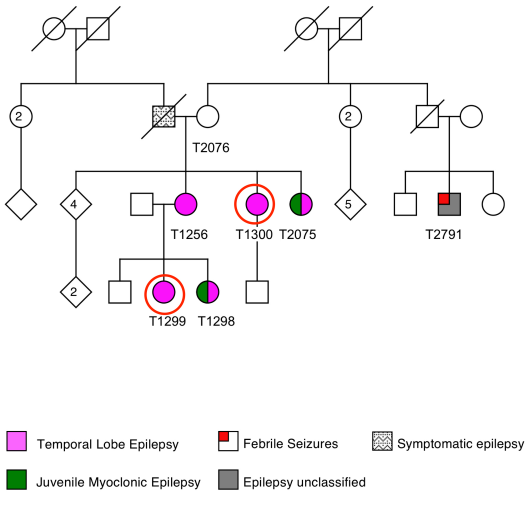
82051

Discovery cohort B
Familial focal epilepsy
"Family B"



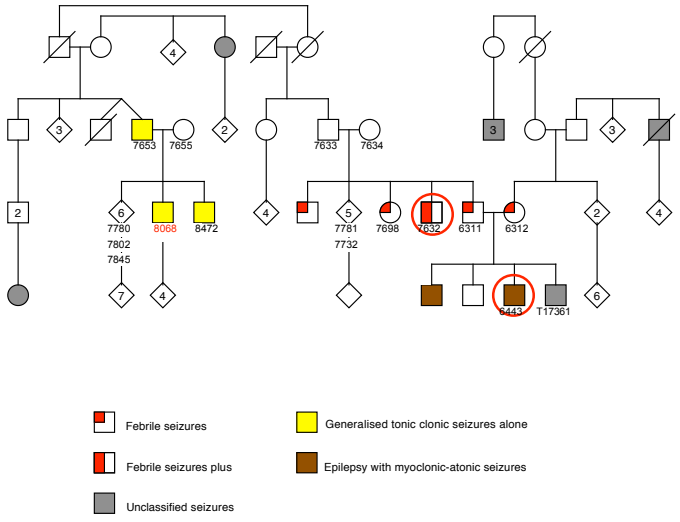
82654

Discovery cohort B
Familial TLE
"Family C"



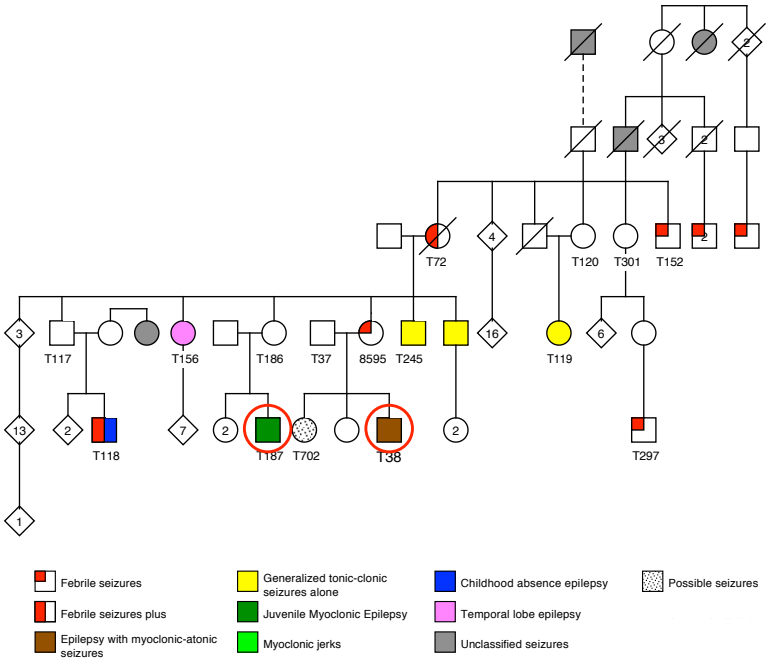
80553

Discovery cohort B
GEFS+
"Family D"



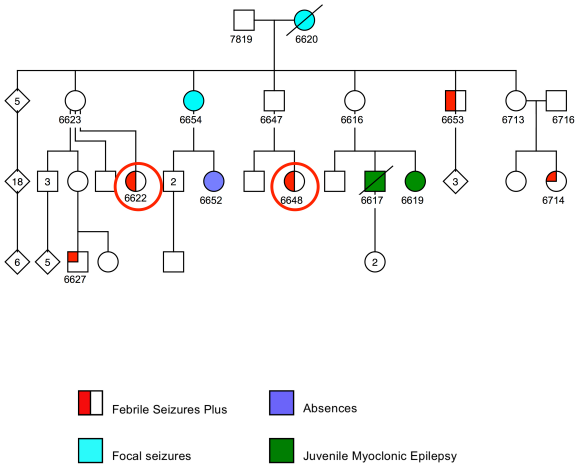
82320

Discovery cohort B
GEFS+/IGE overlap
"Family E"



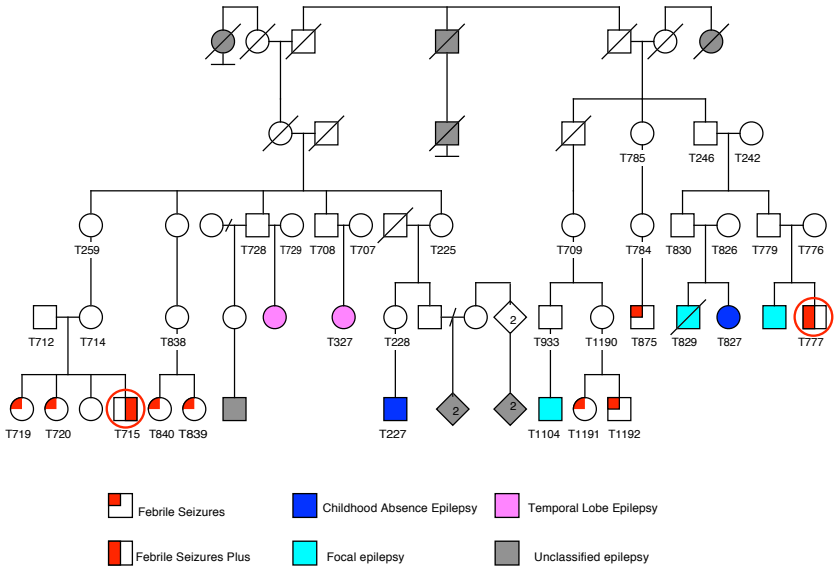
80780

Discovery cohort B
GEFS+/IGE overlap
"Family F"



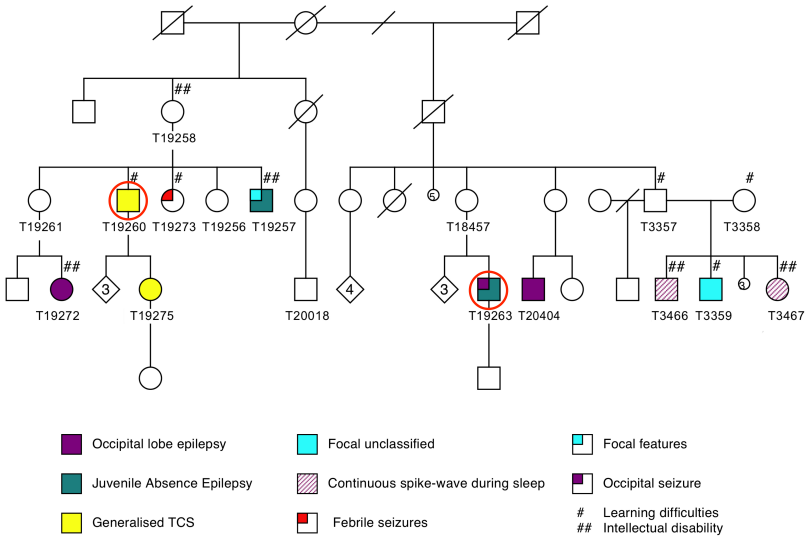
82439

Discovery cohort B
GEFS+
"Family G"



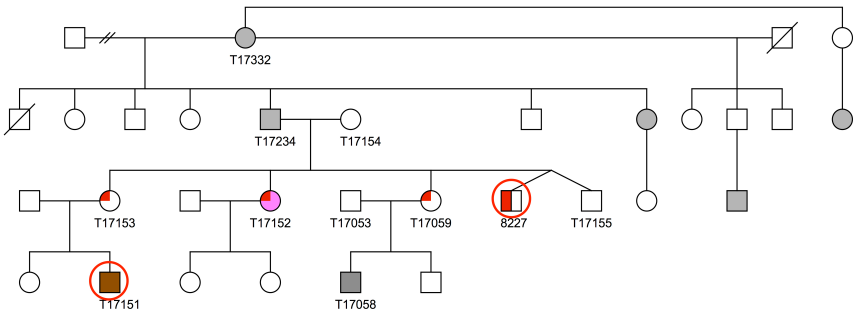
83443

Discovery cohort B
IGE/complex phenotypes
"Family H"



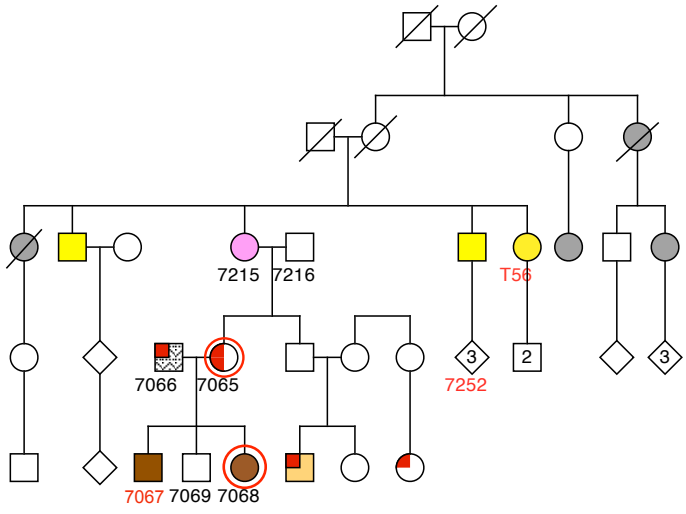
83909

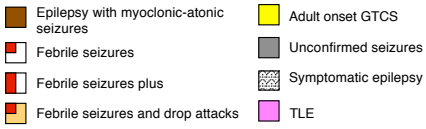
Discovery cohort B
GEFS+
"Family I"



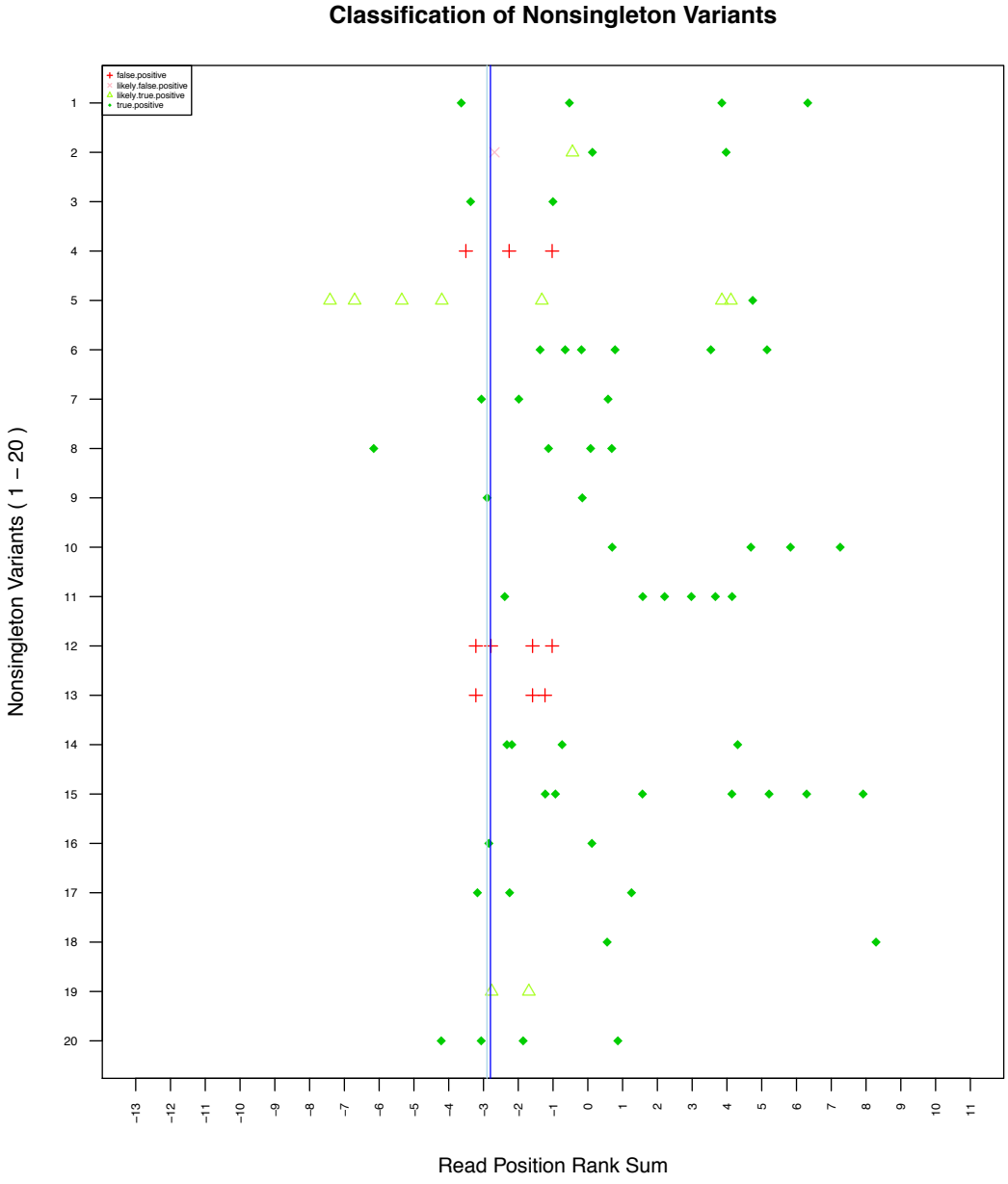
80707

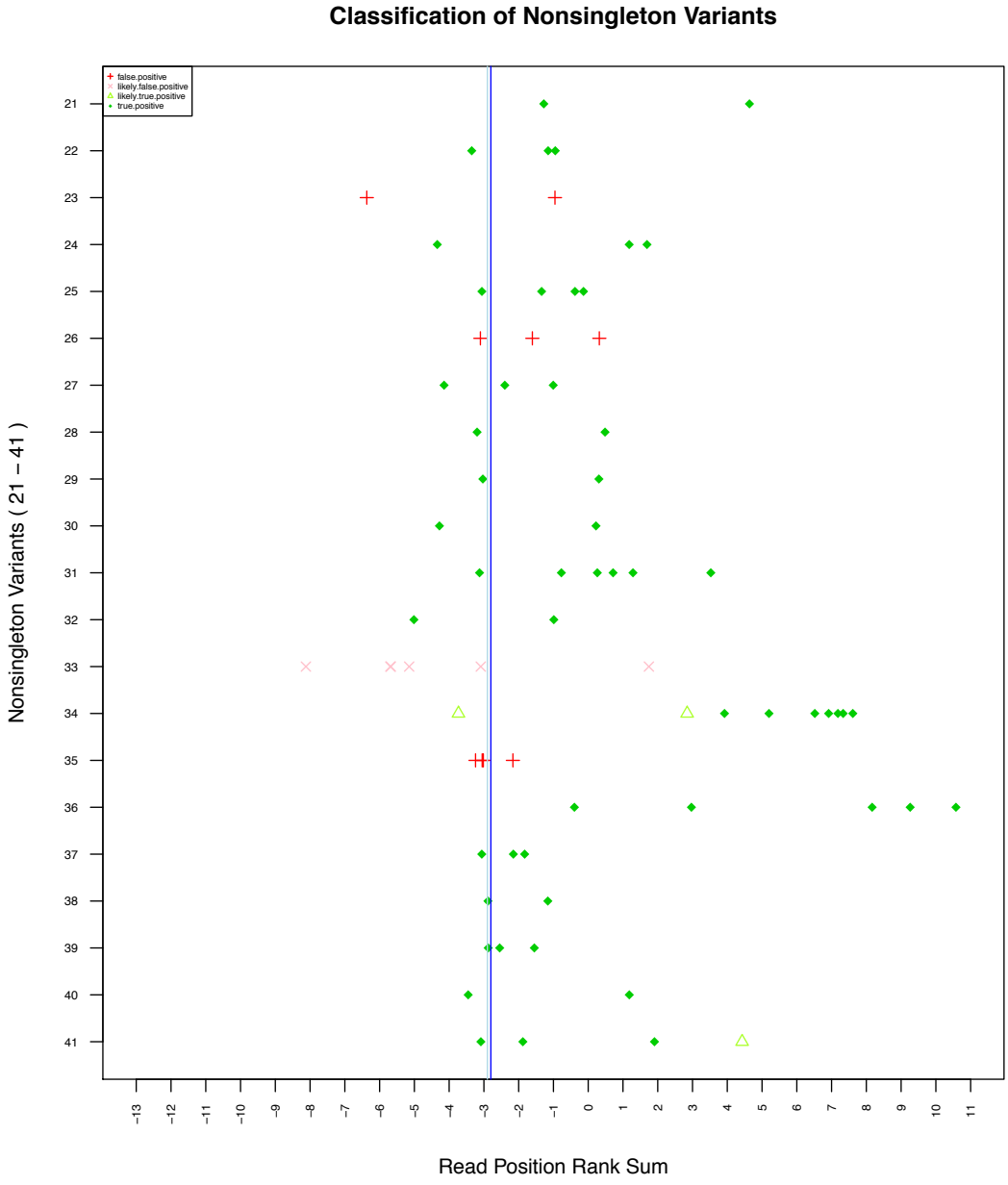
Discovery cohort B
GEFS+
"Family J"

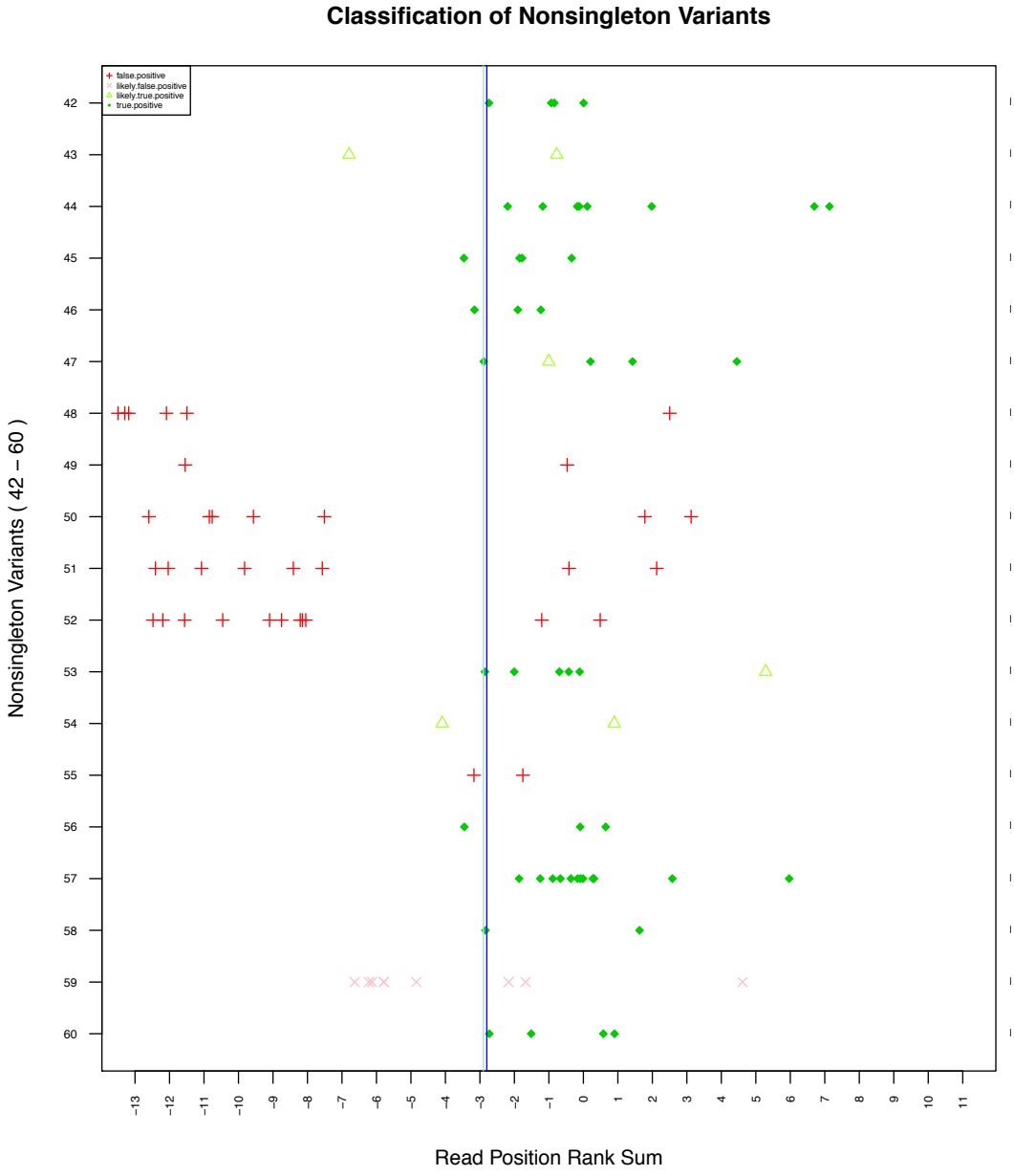


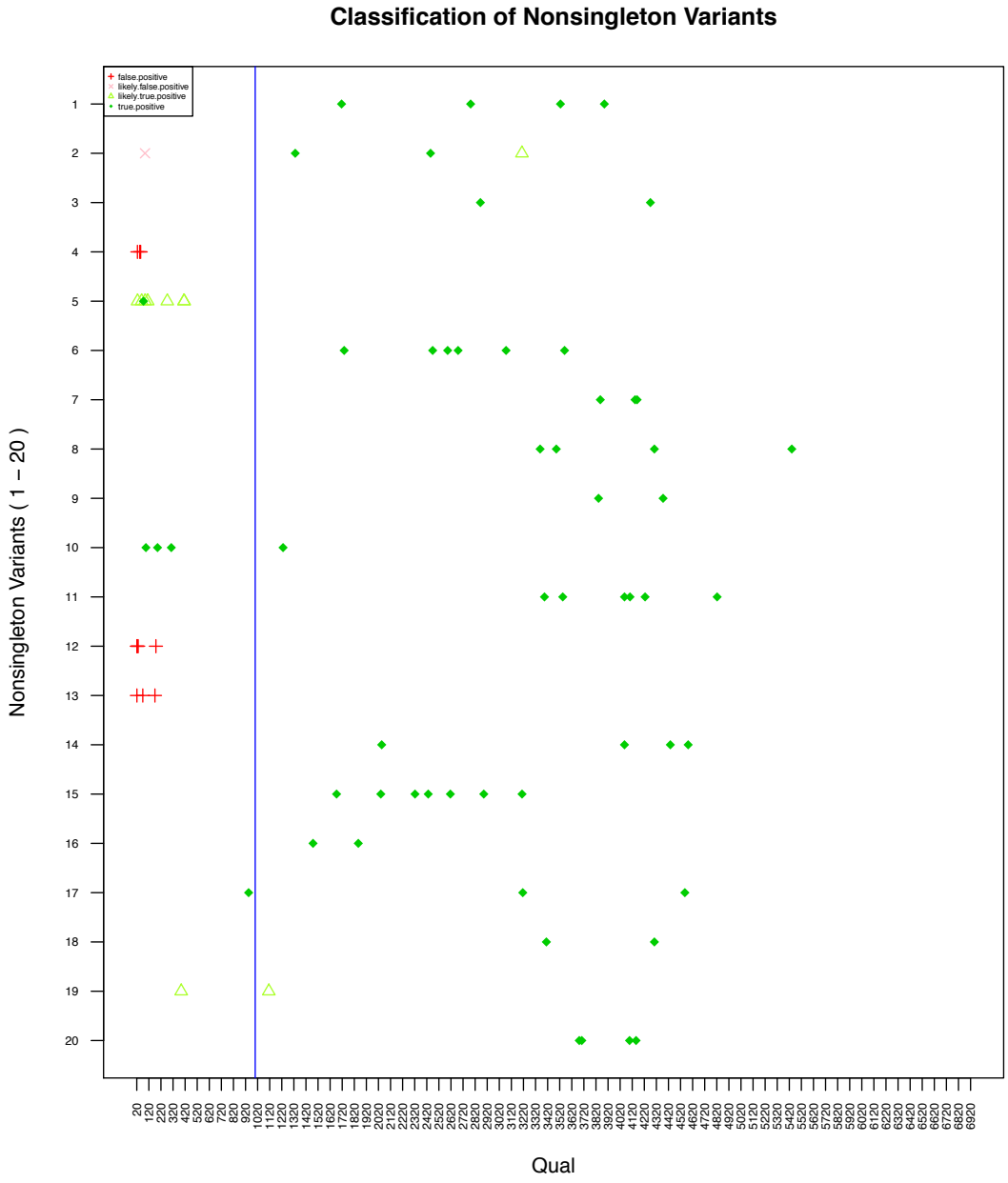


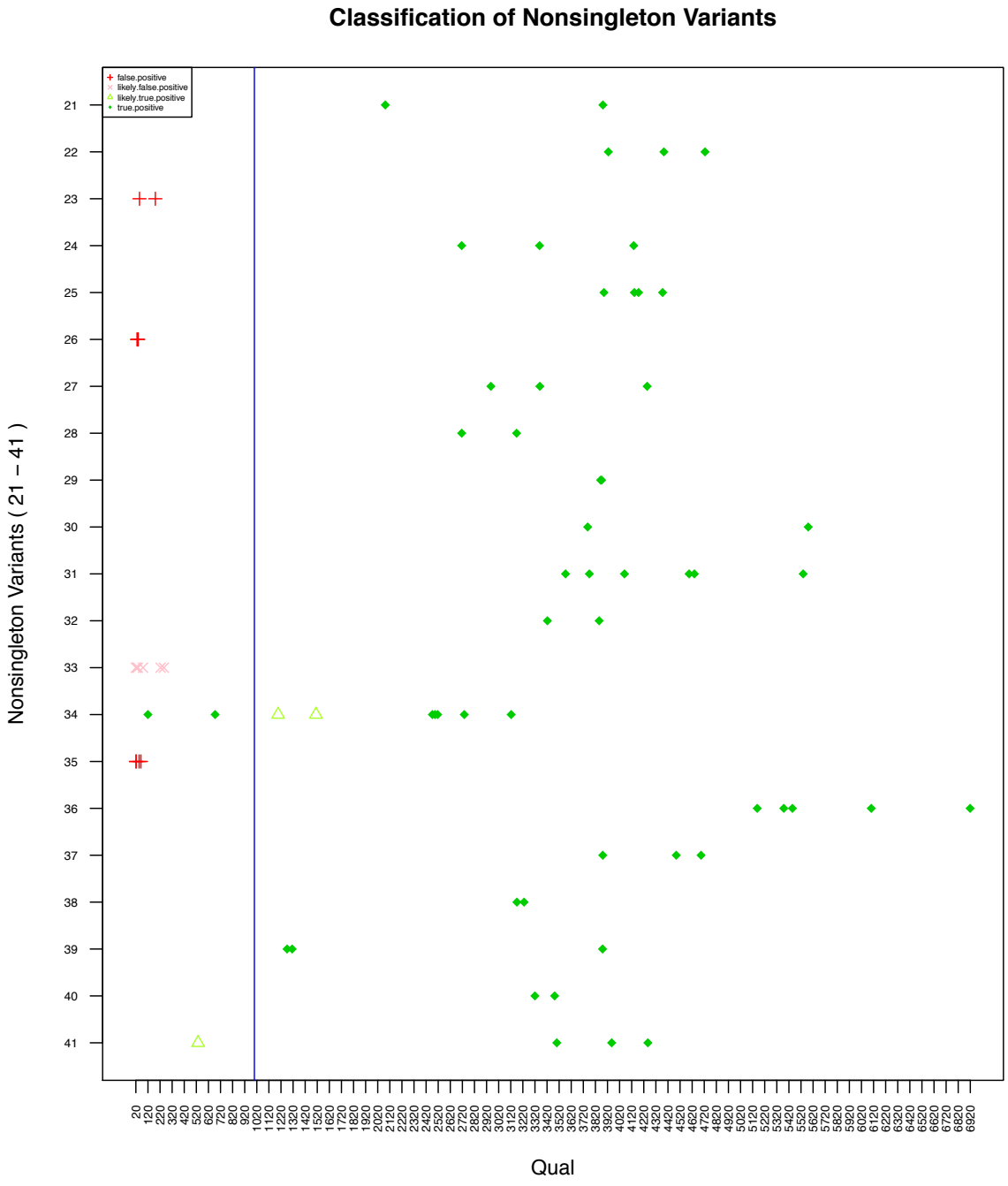
Appendix C: Classification of HaloPlex nonsingleton variants by Read Position Rank Sum and Quality

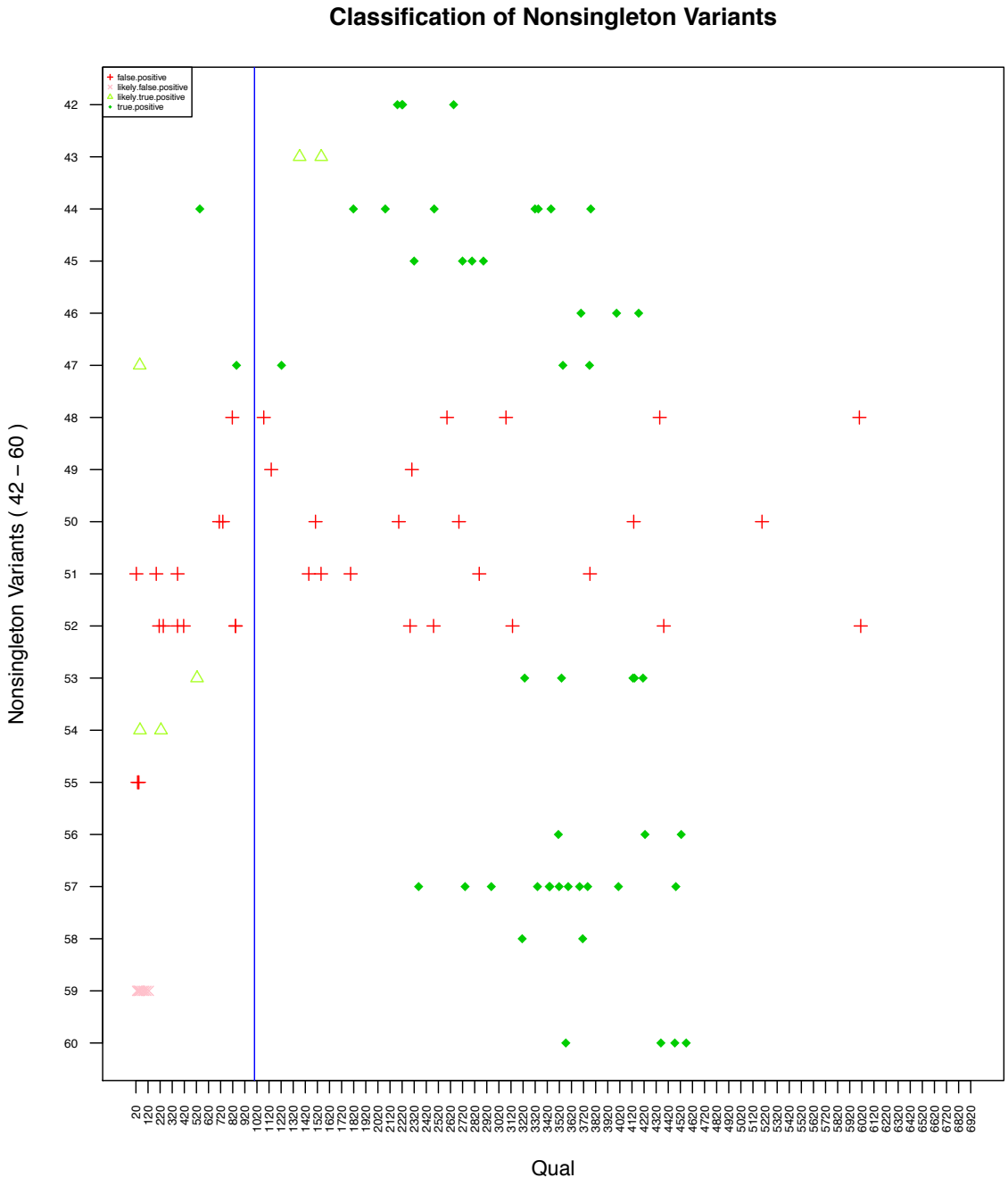



Classification of Nonsingleton Variants
false.positive
likely.false.positive
likely.true.positive
true.positive
Nonsingleton Variants (21 – 41)
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
−13
−12
−11
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
7
8
9
10
11
Read Position Rank Sum


Classification of Nonsingleton Variants
false.positive
likely.false.positive
likely.true.positive
true.positive
Nonsingleton Variants (42 – 60)
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
−13
−12
−11
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
7
8
9
10
11
Read Position Rank Sum


Classification of Nonsingleton Variants
false.positive
likely.false.positive
likely.true.positive
true.positive
Nonsingleton Variants (1 – 20)
Qual


Classification of Nonsingleton Variants
false.positive
likely.false.positive
likely.true.positive
true.positive
Nonsingleton Variants (21 – 41)
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
Qual


Classification of Nonsingleton Variants
false.positive
likely.false.positive
likely.true.positive
true.positive
Nonsingleton Variants (42 – 60)
Qual

References

1. Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, et al. (2010) The characterization of twenty sequenced human genomes. *PLoS Genetics* **6**: e1001111.

2. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11**: 415–425.

3. Beckmann J, Estivill X, Antonarakis S (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Reviews Genetics* **8**: 639–646.

4. Tan NCK, Mulley JC, Berkovic SF (2004) Genetic association studies in epilepsy: "the truth is out there." *Epilepsia* **45**: 1429–1442.

5. McCarthy M, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**: 356–369.

6. Cavalleri GL, Weale ME, Shianna KV, Singh R, Lynch JM, et al. (2007) Multicentre search for genetic susceptibility loci in sporadic epilepsy syndrome and seizure types: a case-control study. *Lancet Neurology* **6**: 970–980.

7. International HapMap Consortium (2003) The International HapMap Project. *Nature* **426**: 789–796.

8. Kruglyak L, Nickerson DA (2001) Variation is the spice of life. *Nature Genetics* **27**: 234–236.

9. Zhu M, Need AC, Han Y, Ge D, Maia JM, et al. (2012) Using ERDS to Infer Copy-Number Variants in High-Coverage Genomes. *The American Journal of Human Genetics* **91**: 408–421.

10. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M (2010) Detecting copy number variation with mated short reads. *Genome Research* **20**: 1613–1622.

11. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, et al. (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Research* **22**: 1525–1532.

12. The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**: 636–640.

13. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, Mcmillin MJ, et al. (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics* **42**: 790–793.

14. Sobreira NLM, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, et al. (2010) Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genetics* **6**: e1000991.

15. Ruzzo EK, Capo-Chichi J-M, Ben-Zeev B, Chitayat D, Mao H, et al. (2013) Deficiency of Asparagine Synthetase Causes Congenital Microcephalyand a Progressive Form of Encephalopathy. *Neuron* **80**: 429–441.

16. Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D (2011) Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic Epidemiology* **35**: 790–799.

17. Hauser WA, Annegers JF, (null) (1996) Descriptive epidemiology of epilepsy: contributions of population-based studies from Rochester, Minnesota. **71**: 576–586.

18. Brodie MJ (2005) Diagnosing and predicting refractory epilepsy. *ACTA Neurologica Scandinavica* **181**: 36–39.

19. Berg AT, Berkovic SF, Brodie MJ, Buchhalter J, Cross JH, et al. (2010) Revised terminology and concepts for organization of seizures and epilepsies: Report of the ILAE Commission on Classification and Terminology, 2005-2009. *Epilepsia* **51**: 676–685.

20. Kjeldsen MJ, Corey LA, Christensen K, Friis ML (2003) Epileptic seizures and syndromes in twins: the importance of genetic factors. *Epilepsy Research* **55**: 137–146.

21. Annegers JF, Hauser WA, (null), (null) (1982) The risks of seizure disorders among relatives of patients with childhood onset epilepsy. *Neurology* **32**: 174–179.

22. Ottman R, Annegers JF, Risch N, Hauser WA, Susser M (1996) Relations of genetic and environmental factors in the etiology of epilepsy. *Annals of Neurology* **39**: 442–449.

23. Ottman R, Lee JH, Risch N, Hauser WA, Susser M (1996) Clinical indicators of genetic susceptibility to epilepsy. *Epilepsia* **37**: 353–361.

24. Jallon P, Loiseau P, Loiseau J (2001) Newly diagnosed unprovoked epileptic seizures: presentation at diagnosis in CAROLE study. *Epilepsia* **42**: 464–475.

25. Berkovic SF, Howell RA, Hay DA, Hooper JL (1998) Epilepsies in twins: genetics of the major epilepsy syndromes. *Annals of Neurology* **43**: 435–445.

26. Berkovic SF, Mulley JC, Scheffer IE, Petrou S (2006) Human epilepsies: interaction of genetic and acquired factors. *Trends in Neurosciences* **29**: 391–397.

27. Ottman R, Hirose S, Jain S, Lerche H, Lopes-Cendes I, et al. (2010) Genetic testing in the epilepsies-Report of the ILAE Genetics Commission. *Epilepsia* **51**: 655–670.

28. Goldstein DB (2009) Common genetic variation and human traits. *The New England Journal of Medicine* **360**: 1696–1698.

29. Iyengar SK, Elston RC (2007) The genetic basis of complex traits: rare variants or "common gene, common disease"? *Methods in Molecular Biology* **376**: 71–84.

30. Heinzen EL, Radtke RA, Urban TJ, Cavalleri GL, Depondt C, et al. (2010) Rare Deletions at 16p13.11 Predispose to a Diverse Spectrum of Sporadic Epilepsy Syndromes. *The American Journal of Human Genetics* **86**: 707–718.

31. Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, et al. (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nature Genetics* **41**: 160–162.

32. Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, et al. (2008) Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *The New England Journal of Medicine* **359**: 1685–1699.

33. Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, et al. (2008) A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics* **40**: 322–328.

34. Stefansson H, Rujescu D, Cichon S, Pietiläinen OPH, Ingason A, et al. (2008) Large recurrent microdeletions associated with schizophrenia. *Nature* **455**: 232–236.

35. Need AC, Ge D, Weale ME, Maia J, Feng S, et al. (2009) A Genome-Wide Investigation of SNPs and CNVs in Schizophrenia. *PLoS Genetics* **5**: e1000373.

36. Steinlein OK, Mulley JC, Propping P, Wallace RH, Phillips HA, et al. (1995) A missense mutation in the neuronal nicotinic acetylcholine receptor alpha 4 subunit is associated with autosomal dominant nocturnal frontal lobe epilepsy. *Nature Genetics* **11**: 201–203.

37. Kalachikov S, Evgrafov O, Ross B, Winawer M, Barker-Cummings C, et al. (2002) Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features. *Nature Genetics* **30**: 335–341.

38. Dibbens LM, de Vries B, Donatello S, Heron SE, Hodgson BL, et al. (2013) Mutations in DEPDC5 cause familial focal epilepsy with variable foci. *Nature Genetics* **45**: 546–551.

39. Ishida S, Picard F, Rudolf G, Noé E, Achaz G, et al. (2013) Mutations of DEPDC5 cause autosomal dominant focal epilepsies. *Nature Genetics* **45**: 552–555.

40. Heron SE, Grinton BE, Kivity S, Afawi Z, Zuberi SM, et al. (2012) PRRT2 Mutations Cause Benign Familial Infantile Epilepsy and Infantile Convulsions with Choreoathetosis Syndrome. *The American Journal of Human Genetics* **90**: 152–160.

41. Carvill GL, Heavin SEAB, Yendle SC, McMahon JM, O'roak BJ, et al. (2013) Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nature Genetics* **45**: 825–830.

42. Epi4K Consortium, Epilepsy Phenome/Genome Project, Allen AS, Berkovic SF, Cossette P, et al. (2013) De novo mutations in epileptic encephalopathies. *Nature* **501**: 217–221.

43. Lemke JR, Riesch E, Scheurenbrand T, Schubach M, Wilhelm C, et al. (2012) Targeted next generation sequencing as a diagnostic tool in epileptic disorders. *Epilepsia* **53**: 1387–1398.

44. Kasperavičiūtė D, Catarino CB, Heinzen EL, Depondt C, Cavalleri GL, et al. (2010) Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study. *Brain* **133**: 2136–2147.

45. Guo Y, Baum LW, Sham PC, Wong V, Ng PW, et al. (2012) Two-stage genome-wide association study identifies variants in CAMSAP1L1 as susceptibility loci for epilepsy in Chinese. *Human Molecular Genetics* **21**: 1184–1189.

46. EPICURE Consortium, EMINet Consortium, Steffens M, Leu C, Ruppert AK, et

al. (2012) Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. *Human Molecular Genetics* **21**: 5359–5372.

47. Escayg A, Goldin AL (2010) Sodium channel SCN1A and epilepsy: mutations and mechanisms. *Epilepsia* **51**: 1650–1658.

48. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. *PLoS Biology* **8**: e1000294.

49. de Vries BBA, Pfundt R, Leisink M, Koolen DA, Vissers LELM, et al. (2005) Diagnostic genome profiling in mental retardation. *The American Journal of Human Genetics* **77**: 606–616.

50. Stone JL, O'donovan MC, Gurling H, Kirov GK, Blackwood DHR, et al. (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**: 237–241.

51. Miller DT, Shen Y, Weiss LA, Korn J, Anselm I, et al. (2009) Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders. *Journal of Medical Genetics* **46**: 242–248.

52. de Kovel CGF, Trucks H, Helbig I, Mefford HC, Baker C, et al. (2010) Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain* **133**: 23–32.

53. Mefford HC, Muhle H, Ostertag P, Spiczak von S, Buysse K, et al. (2010) Genome-Wide Copy Number Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies. *PLoS Genetics* **6**: e1000962.

54. Mefford HC, Yendle SC, Hsu C, Cook J, geraghty E, et al. (2011) Rare copy number variants are an important cause of epileptic encephalopathies. *Annals of Neurology* **70**: 974–985.

55. Pal DK, Pong AW, Chung WK (2010) Genetic evaluation and counseling for epilepsy. *Nature Reviews Neurology* **6**: 445–453.

56. Heinzen EL, Depondt C, Cavalleri GL, Ruzzo EK, Walley NM, et al. (2012) Exome Sequencing Followed by Large-Scale Genotyping Fails to Identify Single Rare Variants of Large Effect in Idiopathic Generalized Epilepsy. *The American Journal of Human Genetics* **91**: 293–302.

57. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, et al. (2010) A de novo paradigm for mental retardation. *Nature Genetics* **42**: 1109–1112.

58. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**: 242–245.

59. O'roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, et al. (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics* **43**: 585–589.

60. Epi4K Consortium (2012) Epi4K: gene discovery in 4,000 genomes. *Epilepsia* **53**: 1457–1467.

61. Kodera H, Kato M, Nord AS, Walsh T, Lee M, et al. (2013) Targeted capture and sequencing for detection of mutations causing early onset epileptic encephalopathy. *Epilepsia* **54**: 1262–1269.

62. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, et al. (2012) De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74**: 285–299.

63. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**: 237–241.

64. Frankel WN (2009) Genetics of complex neurological disease: challenges and opportunities for modeling epilepsy in mice and rats. *Trends in Genetics* **25**: 361–367.

65. Klassen T, Davis C, Goldman A, Burgess D, Chen T, et al. (2011) Exome Sequencing of Ion Channel Genes Reveals Complex Profiles Confounding Personal Risk Assessment in Epilepsy. *Cell* **145**: 1036–1048.

66. Löscher W (2002) Role of Multidrug Transporters in Pharmacoresistance to Antiepileptic Drugs. *Journal of Pharmacology and Experimental Therapeutics* **301**: 7–14.

67. van der Weide J, Steijns LS, van Weelden MJ, de Haan K (2001) The effect of genetic polymorphism of cytochrome P450 CYP2C9 on phenytoin dose requirement. *Pharmacogenetics* **11**: 287–291.

68. Odani A, Hashimoto Y, Otsuki Y, Uwai Y, Hattori H, et al. (1997) Genetic

polymorphism of the CYP2C subfamily and its effect on the pharmacokinetics of phenytoin in Japanese patients with epilepsy. *Clinical Pharmacology and Therapeutics* **62**: 287–292.

69. Aynacioglu AS, Brockmöller J, Bauer S, Sachse C, Güzelbey P, et al. (1999) Frequency of cytochrome P450 CYP2C9 variants in a Turkish population and functional relevance for phenytoin. *British Journal of Clinical Pharmacology* **48**: 409–415.

70. Swen JJ, Nijenhuis M, De Boer A, Grandia L, Maitland-van der Zee AH, et al. (2011) Pharmacogenetics: From Bench to Byte— An Update of Guidelines. *Clinical Pharmacology and Therapeutics* **89**: 662–673.

71. Remy S, Beck H (2006) Molecular and cellular mechanisms of pharmacoresistance in epilepsy. *Brain* **129**: 18–35.

72. Rogawski MA, Johnson MR (2008) Intrinsic severity as a determinant of antiepileptic drug refractoriness. *Epilepsy Currents* **8**: 127–130.

73. Chung W-H, Hung S-I, Hong H-S, Hsih M-S, Yang L-C, et al. (2004) Medical genetics: a marker for Stevens-Johnson syndrome. *Nature* **428**: 486.

74. Leckband SG, Kelsoe JR, Dunnenberger HM, George AL, Tran E, et al. (2013) Clinical Pharmacogenetics Implementation Consortium Guidelines for HLA-B Genotype and Carbamazepine Dosing. *Clinical Pharmacology and Therapeutics* **94**: 324–328.

75. Hung S-I, Chung W-H, Liu Z-S, Chen C-H, Hsih M-S, et al. (2010) Common risk allele in aromatic antiepileptic-drug induced Stevens-Johnson syndrome and toxic epidermal necrolysis in Han Chinese. *Pharmacogenomics* **11**: 349–356.

76. Ozeki T, Mushiroda T, Yowang A, Takahashi A, Kubo M, et al. (2011) Genome-wide association study identifies HLA-A* 3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population. *Human Molecular Genetics* **20**: 1034–1041.

77. McCormack M, Alfirevic A, Bourgeois S, Farrell JJ, Kasperavičiūtė D, et al. (2011) HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. *The New England Journal of Medicine* **364**: 1134–1143.

78. McCormack M, Urban TJ, Shianna KV, Walley N, Pandolfo M, et al. (2012) Genome-wide mapping for clinically relevant predictors of lamotrigine-and

phenytoin-induced hypersensitivity reactions. *Pharmacogenomics* **13**: 399–405.

79. Klepper J, Diefenbach S, Kohlschütter A, Voit T (2004) Effects of the ketogenic diet in the glucose transporter 1 deficiency syndrome. *Prostaglandins, Leukotrienes and Essential Fatty Acids* **70**: 321–327.

80. Mullen SA, Carvill GL, Bellows S, Bayly MA, Berkovic SF, et al. (2013) Copy number variants are frequent in genetic generalized epilepsy with intellectual disability. *Neurology* **81**: 1507–1514.

81. Mahmood S, Ahmad W, Hassan MJ (2011) Autosomal recessive primary microcephaly (MCPH): clinical manifestations, genetic heterogeneity and mutation continuum. *Orphanet Journal of Rare Diseases* **6**: 39.

82. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. *Nucleic Acids Research* **37**: D690–D697.

83. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

84. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

85. Ge D, Ruzzo EK, Shianna KV, He M, Pelak K, et al. (2011) SVA: software for annotating and visualizing sequenced human genomes. *Bioinformatics* **27**: 1998–2000.

86. Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**: 491–498.

87. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**: e164–e164.

88. Scheet P, Stephens M (2006) A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics* **78**: 629–644.

89. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**: 559–575.

90. Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics* **Chapter 2**: Unit2.3.

91. Carlson BR, Lloyd KE, Kruszewski A, Kim I-H, Rodriguiz RM, et al. (2011) WRP/srGAP3 facilitates the initiation of spine development by an inverse F-BAR domain, and its loss impairs long-term memory. *The Journal of Neuroscience* **31**: 2447–2460.

92. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**: 3812–3814.

93. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. *Nature Methods* **7**: 248–249.

94. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**: 297–304.

95. Larsen TM, Boehlein SK, Schuster SM, Richards NGJ, Thoden JB, et al. (1999) Three-Dimensional Structure of Escherichia coli Asparagine Synthetase B: A Short Journey from Substrate to Product. *Biochemistry* **38**: 16146–16157.

96. Chen H, Pan Y-X, Dudenhausen EE, Kilberg MS (2004) Amino acid deprivation induces the transcription rate of the human asparagine synthetase gene through a timed program of expression and promoter binding of nutrient-responsive basic region/leucine zipper transcription factors as well as localized histone acetylation. *The Journal of Biological Chemistry* **279**: 50829–50839.

97. Greco A, Gong S, Ittmann M, Basilico C (1989) Organization and expression of the cell cycle gene, ts11, that encodes asparagine synthetase. *Molecular and Cellular Biology* **9**: 2350.

98. Li BS, Gu LJ, Luo CY, Li WS, Jiang LM, et al. (2006) The downregulation of asparagine synthetase expression can increase the sensitivity of cells resistant to l-asparaginase. *Leukemia* **20**: 2199–2201.

99. Richards NGJ, Kilberg MS (2006) Asparagine Synthetase Chemotherapy. *Annual Review of Biochemistry* **75**: 629–654.

100. Hongo S, Chiyo T, Takeda M (1996) Cloning of cDNA for asparagine synthetase from rat Sertoli cell. *Biochemistry and Molecular Biology International* **38**: 189–196.

101. Visel A, Thaller C, Eichele G (2004) GenePaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucleic Acids Research* **32**: D552–D556.

102. Ayoub AE, Oh S, Xie Y, Leng J, Cotney J, et al. (2011) Transcriptional programs in transient embryonic zones of the cerebral cortex defined by high-resolution mRNA sequencing. *PNAS* **108**: 14950–14955.

103. Bond J, Roberts E, Mochida GH, Hampshire DJ, Scott S, et al. (2002) ASPM is a major determinant of cerebral cortical size. *Nature Genetics* **32**: 316–320.

104. Jackson AP, Eastwood H, Bell SM, Adu J, Toomes C, et al. (2002) Identification of microcephalin, a protein implicated in determining the size of the human brain. *The American Journal of Human Genetics* **71**: 136–142.

105. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25.

106. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

107. Horowitz B, Madras BK, Meister A, Old LJ, Boyes EA, et al. (1968) Asparagine synthetase activity of mouse leukemias. *Science* **160**: 533–535.

108. Adams DJ, van der Weyden L (2008) Contemporary approaches for modifying the mouse genome. *Physiological Genomics* **34**: 225–238.

109. Pulvers JN, Bryk J, Fish JL, Wilsch-Bräuninger M, Arai Y, et al. (2010) Mutations in mouse Aspm (abnormal spindle-like microcephaly associated) cause not only microcephaly but also major defects in the germline. *PNAS* **107**: 16595–16600.

110. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**: 671–675.

111. Rampon C, Tang YP, Goodhouse J, Shimizu E, Kyin M, et al. (2000) Enrichment induces structural changes and recovery from nonspatial memory deficits in CA1 NMDAR1-knockout mice. *Nature Neuroscience* **3**: 238–244.

112. Scholl-Burgi S, Haberlandt E, Heinz-Erian P, Deisenhammer F, Albrecht U, et al. (2008) Amino Acid Cerebrospinal Fluid/Plasma Ratios in Children: Influence of Age, Gender, and Antiepileptic Medication. *Pediatrics* **121**: e920–e926.

113. Nishimura F, Nishihara M, Mori M, Torii K, Takahashi M (1995) Excitability of neurons in the ventromedial nucleus in rat hypothalamic slices: modulation by amino acids at cerebrospinal fluid levels. *Brain Research* **691**: 217–222.

114. Akiyama T, Kobayashi K, Higashikage A, Sato J, Yoshinaga H (2014) CSF/plasma ratios of amino acids: reference data and transports in children. *Brain and Development* **36**: 3–9.

115. Häberle J, Görg B, Rutsch F, Schmidt E, Toutain A, et al. (2005) Congenital glutamine deficiency with glutamine synthetase mutations. *The New England Journal of Medicine* **353**: 1926–1933.

116. van der Crabben SN, Verhoeven-Duif NM, Brilstra EH, Van Maldergem L, Coskun T, et al. (2013) An update on serine deficiency disorders. *Journal of Inherited Metabolic Disease* **36**: 613–619.

117. Häberle J, Shahbeck N, Ibrahim K, Schmitt B, Scheer I, et al. (2012) Glutamine supplementation in a child with inherited GS deficiency improves the clinical status and partially corrects the peripheral and central amino acid imbalance. *Orphanet Journal of Rare Diseases* **7**: 48.

118. Heinzen EL, Swoboda KJ, Hitomi Y, Gurrieri F, Nicole S, et al. (2012) De novo mutations in ATP1A3 cause alternating hemiplegia of childhood. *Nature Genetics* **44**: 1030–1034.

119. Calvo SE, Compton AG, Hershman SG, Lim SC, Lieber DS, et al. (2012) Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. *Science Translational Medicine* **4**: 118ra10.

120. Dixon-Salazar TJ, Silhavy JL, Udpa N, Schroth J, Bielas S, et al. (2012) Exome Sequencing Can Improve Diagnosis and Alter Patient Management. *Science Translational Medicine* **4**: 138ra78–138ra78.

121. Need AC, Shashi V, Hitomi Y, Schoch K, Shianna KV, et al. (2012) Clinical application of exome sequencing in undiagnosed genetic conditions. *Journal of Medical Genetics* **49**: 353–361.

122. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, et al. (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *The New England Journal of Medicine* **369**: 1502–1511.

123. de Ligt J, Willemsen MH, Van Bon BWM, Kleefstra T, Yntema HG, et al. (2012)

Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *The New England Journal of Medicine* **367**: 121003140044006.

124. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB (2013) Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics* **9**: e1003709.

125. O'roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**: 246–250.

126. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, et al. (2012) Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet* **380**: 1674–1682.

127. Barghuti F, Elian K, Gomori JM, Shaag A, Edvardson S, et al. (2008) The unique neuroradiology of complex I deficiency due to NDUFA12L defect. *Molecular Genetics and Metabolism* **94**: 78–82.

128. Dahl HH, Brown GK, Brown RM, Hansen LL, Kerr DS, et al. (1992) Mutations and polymorphisms in the pyruvate dehydrogenase E1 alpha gene. *Human Mutation* **1**: 97–102.

129. Anikster Y, Shaag A, Joseph A, Mandel H, Ben-Zeev B, et al. (1996) Glutaric aciduria type I in the Arab and Jewish communities in Israel. *The American Journal of Human Genetics* **59**: 1012–1018.

130. Korman SH, Jakobs C, Darmin PS, Gutman A, van der Knaap MS, et al. (2007) Glutaric aciduria type 1: Clinical, biochemical and molecular findings in patients from Israel. *European Journal of Paediatric Neurology* **11**: 81–89.

131. Ferreira H, Seppala R, Pinto R, Huizing M, Martins E, et al. (1999) Sialuria in a Portuguese girl: clinical, biochemical, and molecular characteristics. *Molecular Genetics and Metabolism* **67**: 131–137.

132. Hoischen A, Van Bon BWM, Rodríguez-Santiago B, Gilissen C, Vissers LELM, et al. (2011) De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nature Genetics* **43**: 729–731.

133. Heron SE, Smith KR, Bahlo M, Nobili L, Kahana E, et al. (2012) Missense mutations in the sodium-gated potassium channel gene KCNT1 cause severe autosomal dominant nocturnal frontal lobe epilepsy. *Nature Genetics* **44**: 1188–

1190.

134. Wada T, Kubota T, Fukushima Y, Saitoh S (2000) Molecular genetic study of Japanese patients with X-linked α-thalassemia/mental retardation syndrome (ATR-X). *American Journal of Medical Genetics* **94**: 242–248.

135. Hood RL, Lines MA, Nikkel SM, Schwartzentruber J, Beaulieu C, et al. (2012) Mutations in SRCAP, Encoding SNF2-Related CREBBP Activator Protein, Cause Floating-Harbor Syndrome. *The American Journal of Human Genetics* **90**: 308–313.

136. Nakamura K, Kodera H, Akita T, Shiina M, Kato M, et al. (2013) De Novo Mutations in GNAO1, Encoding a G alpha o Subunit of Heterotrimeric G Proteins, Cause Epileptic Encephalopathy. *The American Journal of Human Genetics* **93**: 496–505.

137. Bainbridge MN, Hu H, Muzny DM, Musante L, Lupski JR, et al. (2013) De novo truncating mutations in ASXL3 are associated with a novel clinical phenotype with similarities to Bohring-Opitz syndrome. *Genome Medicine* **5**: 11.

138. Megarbane A, Pangrazio A, Villa A, Chouery E, Maarawi J, et al. (2013) Homozygous stop mutation in the SNX10 gene in a consanguineous Iraqi boy with osteopetrosis and corpus callosum hypoplasia. *European Journal of Medical Genetics* **56**: 32–35.

139. Aker M, Rouvinski A, Hashavia S, Ta-Shma A, Shaag A, et al. (2012) An SNX10 mutation causes malignant osteopetrosis of infancy. *Journal of Medical Genetics* **49**: 221–226.

140. Oz-Levi D, Ben-Zeev B, Ruzzo EK, Hitomi Y, Gelman A, et al. (2012) Mutation in TECPR2 reveals a role for autophagy in hereditary spastic paraparesis. *The American Journal of Human Genetics* **91**: 1065–1072.

141. Ng D, Thakker N, Corcoran CM, Donnai D, Perveen R, et al. (2004) Oculofaciocardiodental and Lenz microphthalmia syndromes result from distinct classes of mutations in BCOR. *Nature Genetics* **36**: 411–416.

142. Agamy O, Ben-Zeev B, Lev D, Marcus B, Fine D, et al. (2010) Mutations disrupting selenocysteine formation cause progressive cerebello-cerebral atrophy. *The American Journal of Human Genetics* **87**: 538–544.

143. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. *PNAS* **104**: 8685–8690.

144. Dedek K, Fusco L, Teloy N, Steinlein OK (2003) Neonatal convulsions and epileptic encephalopathy in an Italian family with a missense mutation in the fifth transmembrane region of KCNQ2. *Epilepsy Research* **54**: 21–27.

145. Cacciagli P, Haddad M-R, Mignon-Ravix C, El-Waly B, Moncla A, et al. (2010) Disruption of the ATP8A2 gene in a patient with a t(10;13) de novo balanced translocation and a severe neurological phenotype. *European Journal of Human Genetics* **18**: 1360–1363.

146. Barcia G, Fleming MR, Deligniere A, Gazula V-R, Brown MR, et al. (2012) De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. *Nature Genetics* **44**: 1255–1259.

147. Kaufmann R, Straussberg R, Mandel H, Fattal-Valevski A, Ben-Zeev B, et al. (2010) Infantile cerebral and cerebellar atrophy is associated with a mutation in the MED17 subunit of the transcription preinitiation mediator complex. *The American Journal of Human Genetics* **87**: 667–670.

148. Bouzidi M, Tricaud N, Giraud P, Kordeli E, Caillol G, et al. (2002) Interaction of the Nav1.2a subunit of the voltage-dependent sodium channel with nodal ankyrinG. *The Journal of Biological Chemistry* **277**: 28996–29004.

149. Malhotra JD, Koopmann MC, Kazen-Gillespie KA, Fettman N, Hortsch M, et al. (2002) Structural requirements for interaction of sodium channel beta 1 subunits with ankyrin. *The Journal of Biological Chemistry* **277**: 26681–26688.

150. Ogiwara I, Ito K, Sawaishi Y, Osaka H, Mazaki E, et al. (2009) De novo mutations of voltage-gated sodium channel II gene SCN2A in intractable epilepsies. *Neurology* **73**: 1046–1053.

151. Wallace RH, Wang DW, Singh R, Scheffer IE, George AL, et al. (1998) Febrile seizures and generalized epilepsy associated with a mutation in the Na+-channel beta1 subunit gene SCN1B. *Nature Genetics* **19**: 366–370.

152. Iqbal Z, Vandeweyer G, van der Voet M, Waryah AM, Zahoor MY, et al. (2013) Homozygous and heterozygous disruptions of ANK3: at the crossroads of neurodevelopmental and psychiatric disorders. *Human Molecular Genetics* **22**: 1960–1970.

153. Shashi V, McConkie-Rosell A, Rosell B, Schoch K, Vellore K, et al. (2013) The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genetics in*

Medicine **16**: 176–182.

154. Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, et al. (2012) Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. *Science Translational Medicine* **4**: 154ra135–154ra135.

155. The EPGP Collaborative, McGovern K, Stillman N, McKenna K, Mays V, et al. (2013) The Epilepsy Phenome/Genome Project. *Clinical Trials* **10**: 568–586.

156. Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies. *The American Journal of Human Genetics* **80**: 727–739.

157. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**: 471–475.

158. Carvill GL, Weckhuysen S, McMahon JM, Hartmann C, Møller RS, et al. (2014) GABRA1 and STXBP1: Novel genetic causes of Dravet syndrome. *Neurology* **82**: 1245–1253.

159. Lemke JR, Hendrickx R, Geider K, Laube B, Schwake M, et al. (2014) GRIN2B mutations in West syndrome and intellectual disability with focal epilepsy. *Annals of Neurology* **75**: 147–154.

160. Veeramah KR, O'Brien JE, Meisler MH, Cheng X, Dib-Hajj SD, et al. (2012) De novo pathogenic SCN8A mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP. *The American Journal of Human Genetics* **90**: 502–510.

161. Suls A, Jaehn JA, Kecskés A, Weber Y, Weckhuysen S, et al. (2013) De Novo Loss-of-Function Mutations in CHD2 Cause a Fever-Sensitive Myoclonic Epileptic Encephalopathy Sharing Features with Dravet Syndrome. *The American Journal of Human Genetics* **93**: 967–975.

162. Endele S, Rosenberger G, Geider K, Popp B, Tamer C, et al. (2010) Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nature Publishing Group* **42**: 1021–1026.

163. Lesca G, Rudolf G, Bruneau N, Lozovaya N, Labalme A, et al. (2013) GRIN2A mutations in acquired epileptic aphasia and related childhood focal epilepsies

and encephalopathies with speech and language dysfunction. *Nature Genetics* **45**: 1061–1066.

164. Kodera H, Nakamura K, Osaka H, Maegaki Y, Haginoya K, et al. (2013) De Novo Mutations in SLC35A2Encoding a UDP-Galactose Transporter Cause Early-Onset Epileptic Encephalopathy. *Human Mutation* **34**: 1708–1714.

165. Hamdan FF, Gauthier J, Spiegelman D, Noreau A, Yang Y, et al. (2009) Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. *The New England Journal of Medicine* **360**: 599–605.

166. Hamdan FF, Daoud H, Piton A, Gauthier J, Dobrzeniecka S, et al. (2011) De novo SYNGAP1 mutations in nonsyndromic intellectual disability and autism. *Biological Psychiatry* **69**: 898–901.

167. Wang K, Li M, (null), (null), (null), et al. (2007) PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**: 1665–1674.

168. Need AC, Mcevoy JP, Gennarelli M, Heinzen EL, Ge D, et al. (2012) Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *The American Journal of Human Genetics* **91**: 303–312.

169. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**: 904–909.

170. Dibbens LM, Mullen S, Helbig I, Mefford HC, Bayly MA, et al. (2009) Familial and sporadic 15q13.3 microdeletions in idiopathic generalized epilepsy: precedent for disorders with complex inheritance. *Human Molecular Genetics* **18**: 3626–3631.

171. Morrow EM, Yoo S-Y, Flavell SW, Kim T-K, Lin Y, et al. (2008) Identifying Autism Loci and Genes by Tracing Recent Shared Ancestry. *Science* **321**: 218–223.

172. Xu B, Roos JL, Levy S, Van Rensburg EJ, Gogos JA, et al. (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. *Nature Genetics* **40**: 880–885.

173. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin--rapid analysis

of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**: 97–101.

174. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83**: 311–321.

175. Guo Q, Xie J, Dang CV, Liu ET, Bishop JM (1998) Identification of a large Myc-binding protein that contains RCC1-like repeats. *PNAS* **95**: 9172–9177.

176. Garbarini N, Delpire E (2008) The RCC1 domain of protein associated with Myc (PAM) interacts with and regulates KCC2. *Cellular Physiology and Biochemistry* **22**: 31–44.

177. Wan HI, DiAntonio A, Fetter RD, Bergstrom K, Strauss R, et al. (2000) Highwire regulates synaptic growth in Drosophila. *Neuron* **26**: 313–329.

178. Sampathkumar P, Ozyurt SA, Miller SA, Bain KT, Rutter ME, et al. (2010) Structures of PHR Domains from Mus musculus Phr1 (Mycbp2) Explain the Loss-of-Function Mutation (Gly1092→Glu) of the C. elegans Ortholog RPM-1. *Journal of Molecular Biology* **397**: 883–892.

179. Sakai Y, Shaw CA, Dawson BC, Dugas DV, Al-Mohtaseb Z, et al. (2011) Protein Interactome Reveals Converging Molecular Pathways Among Autism Disorders. *Science Translational Medicine* **3**: 86ra49–86ra49.

180. Ruzzo EK, Pappas AL, Goldstein DB (2012) Modifier genetics in neuropsychiatric disease: challenges and opportunities. *Genome Biology* **13**: 150.

181. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**: 272–276.

182. Ku CS, Naidoo N, Pawitan Y (2011) Revisiting Mendelian disorders through exome sequencing. *Human Genetics* **129**: 351–370.

183. Gilissen C, Hoischen A, Brunner HG, Veltman JA (2011) Unlocking Mendelian disease using exome sequencing. *Genome Biology* **12**: 228.

184. Paşca SP, Portmann T, Voineagu I, Yazawa M, Shcheglovitov A, et al. (2011) Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome. *Nature Medicine* **17**: 1657–1662.

185. Ramsey BW, Davies J, McElvaney NG, Tullis E, Bell SC, et al. (2011) A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. *The New England Journal of Medicine* **365**: 1663–1672.

186. Cho SW, Kim S, Kim JM, Kim J-S (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature Biotechnology* **31**: 230–232.

187. Boone C, Bussey H, Andrews BJ (2007) Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics* **8**: 437–449.

188. Antonarakis SE, Beckmann JS (2006) Mendelian disorders deserve more attention. *Nature Reviews Genetics* **7**: 277–282.

189. Albers CA, Paul DS, Schulze H, Freson K, Stephens JC, et al. (2012) Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nature Genetics* **44**: 435–439.

190. Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, et al. (2010) A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nature Genetics* **42**: 203–209.

191. Bilguvar K, Öztürk AK, Louvi A, Kwan KY, Choi M, et al. (2010) Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467**: 207–210.

192. Nicholas AK, Khurshid M, Désir J, Carvalho OP, Cox JJ, et al. (2010) WDR62 is associated with the spindle pole and is mutated in human microcephaly. *Nature Genetics* **42**: 1010–1014.

193. Guernsey DL, Jiang H, Hussin J, Arnold M, Bouyakdan K, et al. (2010) Mutations in Centrosomal Protein CEP152 in Primary Microcephaly Families Linked to MCPH4. *The American Journal of Human Genetics* **87**: 40–51.

194. Kaindl AM, Passemard S, Kumar P, Kraemer N, Issa L, et al. (2010) Many roads lead to primary autosomal recessive microcephaly. *Progress in Neurobiology* **90**: 363–383.

195. Cox J, Jackson AP, Bond J, Woods CG (2006) What primary microcephaly can tell us about brain growth. *Trends in Molecular Medicine* **12**: 358–366.

196. Kelley RI, Robinson D, Puffenberger EG, Strauss KA, Morton DH (2002) Amish

lethal microcephaly: A new metabolic disorder with severe congenital microcephaly and 2-ketoglutaric aciduria. *American Journal of Medical Genetics* **112**: 318–326.

197. Siu VM, Ratko S, Prasad AN, Prasad C, Rupar CA (2010) Amish microcephaly: Long-term survival and biochemical characterization. *American Journal of Medical Genetics* **152A**: 1747–1751.

198. Hart CE, Race V, Achouri Y, Wiame E, Sharrard M, et al. (2007) Phosphoserine aminotransferase deficiency: a novel disorder of the serine biosynthesis pathway. *The American Journal of Human Genetics* **80**: 931–937.

Biography

Elizabeth Kathryn Ruzzo was born on August 15th, 1982 in Seattle, Washington. She is the youngest daughter of Walter Lawrence and Barbara Jean Ruzzo and sister of Emily Jane Ruzzo. Elizabeth attended the University of Washington in Seattle, Washington and graduated in June 2005 with a B.S. in Cell and Molecular Biology and minors in Chemistry and Anthropology. As an undergraduate, she conducted research on understanding the ecological consequences of declining biodiversity in the laboratory of Dr. Shahid Naeem. After graduation, she worked at Apex Learning to develop digital curriculum for high school students. Here she managed teams of content writers and artists to develop 10th grade biology and 11th grade chemistry courses. In 2007 she went back to research, working in Phillip F. Chance's lab at the University of Washington and investigating the genetic basis of a rare neurological disorder: hereditary neuralgic amyotrophy (HNA). She then entered the University Program in Genetics and Genomics at Duke University in Durham, North Carolina in August of 2008.

Fellowships, memberships, and academic honors

Jo Rae Wright Fellowship for Outstanding Women in Science, Duke University (2012)

Predoctoral Research and Training Fellowship, Epilepsy Foundation (2012)

Ray J. Tysor Graduate Fellowship, Duke University (2010)

Poster Presentation Award, Duke University: Neurobiology Retreat (2011)

Poster Presentation Award, American Epilepsy Society (2010)

Duke University Graduate School Board of Visitors Member (2012-present)

American Epilepsy Society Member

American Society of Human Genetics Member

Peer-reviewed publications

1. **Ruzzo EK**, Capo-Chichi JM, Ben-Zeev B, Chitayat D, Mao H, Pappas AL, Hitomi Y, Lu Y, Hamdan FF, Pelak K, Reznik-Wolf H, Bar-Joseph I, Oz-Levi D, Lev D, Lerman-Sagie T, Leshinsky-Silver E, Anikster Y, Ben-Asher E, Olender T, Colleaux L, Décarie J, Blaser S, Banwell B, Joshi RB, He X, Patry L, Silver RJ, Dobrzeniecka S, Islam MS, Hasnat A, Aryal DK, Rodriguiz RM, Wetsel WC, McNamara JO, Rouleau GA, Silver DL, Lancet D, Pras E, Mitchell GA, Michaud JL, Goldstein DB. Deficiency of asparagine synthetase causes congenital microcephaly and a progressive form of encephalopathy. *Neuron*. 16 October 2013.
2. Epi4K Consortium; Epilepsy Phenome/Genome Project (Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, Epstein MP, Glauser T, Goldstein DB, Han Y, Heinzen EL, Hitomi Y, Howell KB, Johnson MR, Kuzniecky R, Lowenstein DH, Lu YF, Madou MR, Marson AG, Mefford HC, Esmaeeli Nieh S, O'Brien TJ, Ottman R, Petrovski S, Poduri A, **Ruzzo EK**, Scheffer IE, Sherr EH, Yuskaitis CJ, Abou-Khalil B, Alldredge BK, Bautista JF, Berkovic SF, Boro A, Cascino GD, Consalvo D, Crumrine P, Devinsky O, Dlugos D, Epstein MP, Fiol M, Fountain NB, French J, Friedman D, Geller EB, Glauser T, Glynn S, Haut SR, Hayward J, Helmers SL, Joshi S, Kanner A, Kirsch HE, Knowlton RC, Kossoff EH, Kuperman R, Kuzniecky R, Lowenstein DH, McGuire SM, Motika PV, Novotny EJ, Ottman R, Paolicchi JM, Parent JM, Park K, Poduri A, Scheffer IE, Shellhaas RA, Sherr EH, Shih JJ, Singh R, Sirven J, Smith MC, Sullivan J, Lin Thio L, Venkat A, Vining EP, Von Allmen GK, Weisenberg JL, Widdess-Walsh P, Winawer MR). De novo mutations in epileptic encephalopathies. *Nature*. 12 September 2013, PubMed ID: 23934111.
3. Oz-Levi D*, Ben-Zeev B*, **Ruzzo EK**, Hitomi Y, Gelman, A, Elazar, Z, Pelak, K, Anikster, Y, Reznik-Wolf, H, Bar-Joseph, I, Olender, T, Alkelai, A, Ben-Asher, E, Ge, D, Shianna, KV Goldstein, DB, Pras, E, and Lancet, D. Mutation in TECPR2 reveals a role for autophagy in hereditary spastic paraparesis. *American Journal of Human Genetics*. 7 December 2012, PubMed ID: 23176824.

4. Zhu M, Need AC, Han Y, Ge D, Maia JM, Zhu Q, Heinzen EL, Cirulli ET, Pelak K, He M, **Ruzzo EK**, Gumbs C, Singh A, Feng S, Shianna KV, Goldstein DB. Using ERDS to infer copy-number variants in high-coverage genomes. *American Journal of Human Genetics*. 7 September 2012, PubMed ID: 22939633.
5. Heinzen EL, Depondt C, Cavalleri G, **Ruzzo EK**, Walley NM, Need AC, Ge D, He M, Cirulli ET, Zhao Q, Cronin KD, Gumbs CE, Campbell CR, Hong LK, Maia JM, Shianna KV, McCormack M, Radtke RA, Mikati MA, Gallentine WB, Husain AM, Sinha SR, Chinthapalli K, Purnam RS, McNamara JO, Ottman R, Sisodiya SM, Delanty N, Goldstein DB. Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. *American Journal of Human Genetics*. 10 August 2012, PubMed ID: 22863189.
6. **Ruzzo EK**, Pappas AL, Goldstein DB. Modifier genetics in neuropsychiatric disease: Challenges and opportunities. *Genome Biology*. 28 March 2012, PubMed ID: 22458452.
7. Ge D*, **Ruzzo EK***, Shianna KV, He M, Pelak K, Heinzen EL, Need AC, Cirulli ET, Maia JM, Dickson SP, Zhu M, Singh A, Allen AS, Goldstein DB. SVA: Software for Annotating and Visualizing Sequenced Human Genomes. *Bioinformatics*. 15 July 2011, PubMed ID: 21624899.
8. Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, Cirulli ET, Fellay J, Dickson SP, Gumbs CE, Heinzen EL, Need AC, **Ruzzo EK**, Singh A, Campbell CR, Hong LK, Lornsen KA, McKenzie AM, Sobreira NL, Hoover-Fong JE, Milner JD, Ottman R, Haynes BF, Goedert JJ, Goldstein DB. The characterization of twenty sequenced human genomes. *PLoS Genetics*. 9 September 2010, PubMed ID: 20838461.
9. Collie AM, Landsverk ML, **Ruzzo E**, Mefford HC, Buysse K, Adkins JR, Knutzen DM, Barnett K, Brown RH Jr, Parry GJ, Yum SW, Simpson DA, Olney RK, Chinnery PF, Eichler EE, Chance PF, Hannibal MC. Non-recurrent SEPT9 duplications cause hereditary neuralgic amyotrophy. *Journal of Medical Genetics*. September 2010, PubMed ID: 19939853.
10. Hannibal MC, **Ruzzo EK**, Miller LR, Betz B, Buchan JG, Knutzen DM, Barnett K, Landsverk ML, Brice A, LeGuern E, Bedford HM, Worrall BB, Lovitt S, Appel SH, Andermann E, Bird TD, Chance PF. SEPT9 gene sequencing analysis reveals recurrent mutations in hereditary neuralgic amyotrophy. *Neurology*. 19 May 2009, PubMed ID: 19451530.
11. Landsverk ML*, **Ruzzo EK***, Mefford HC, Buysse K, Buchan JG, Eichler EE, Petty EM, Peterson EA, Knutzen DM, Barnett K, Farlow MR, Caress J, Parry GJ, Quan D, Gardner KL, Hong M, Simmons Z, Bird TD, Chance PF, Hannibal MC. Duplication within the SEPT9 gene associated with a founder effect in North American families with hereditary neuralgic amyotrophy. *Human Molecular Genetics*, 1 April 2009, PubMed ID: 19139049.

Platform presentations

1. "Analysis of existing multiplex families: Lessons for Project 2". Epi4K annual meeting. Washington D.C. (December 2013)
2. "Investigating the genetic etiology of familial epilepsies using next- generation sequencing". American Society of Human Genetics Annual Meeting. San Francisco, CA (November 2012)
3. "Multiplex epilepsy families from whole-genome to targeted capture" EPIGEN annual meeting. London, England (October 2012)
4. "Analysis of existing multiplex families: Lessons for Project 2" Epi4K grant kick-off meeting. Durham, NC (December 2011)
5. "Interpreting familial NGS data for the identification of genetic variants influencing epilepsy susceptibility" EPIGEN annual meeting. Slane, Ireland (October 2011)
6. "An overview of studied diseases" MEDIN collaboration meeting, Weizmann Institute of Science. Rehovot, Israel (June 2011)
7. "Exome sequencing identifies a recessive mutation as a cause of progressive microcephaly and brain atrophy" University Program in Genetics and Genomics annual retreat. Asheville, NC (May 2011)
8. "Whole-exome sequencing identifies a recessive mutation as a cause of progressive microcephaly and brain atrophy" Duke University's Department of Molecular Genetics and Microbiology monthly research meeting. Durham, NC (February 2011)
9. "Whole-genome sequencing in multiplex epilepsy families: an approach to identify rare susceptibility variants" EPIGEN annual meeting. Istanbul, Turkey (September 2010)
10. "Examination of a cohort of families to identify rare variants that contribute to epilepsy" EPIGEN annual meeting. Brussels, Belgium (September 2009)

Poster presentations

1. Ruzzo EK, Heinzen EL, Wedel R, Scheffer IE, Berkovic SF, Ottman R, and Goldstein DB. "The genetic etiology of familial epilepsies: from whole-genome to targeted gene sequencing". Gordon Research Conference: Genetics and Genomics. Smithfield, RI (July 2013)
2. Ruzzo EK, Heinzen EL, Wedel R, Shianna KV, Ge D, Ottman R, and Goldstein DB. "Whole-genome sequencing within multiplex epilepsy families: identification of genetic variants influencing epilepsy susceptibility". *Coming Together on Epilepsy* Genetics: From Human to Model Organisms and Back. Bar Harbor, ME (October 2011)

3. Ruzzo EK, Heinzen EL, Wedel R, Shianna KV, Ge D, Ottman R, and Goldstein DB. "Interpreting familial whole-genome sequencing data for the identification of genetic variants influencing epilepsy susceptibility". American Society of Human Genetics Annual Meeting. Montreal, Canada (October 2011)
4. Ruzzo EK, Heinzen EL, Poduri A, Wedel R, Ge D, Shianna KV, Ottman R, and Goldstein DB. "Whole-genome sequencing in multiplex epilepsy families: an approach to identify rare susceptibility variants". American Epilepsy Society Annual Meeting. San Antonio, TX (December 2010)
5. Ruzzo EK, Heinzen EL, Ottman R, Goldstein DB. "A familial whole-genome sequencing approach to identify genetic variants influencing epilepsy susceptibility". American Society of Human Genetics Annual Meeting. Washington, D.C. (November 2010)